Under the editorship of

M. F. NIMKOFF

The Florida State University

Statistical Reasoning in Sociology

John H. Mueller
Karl F. Schuessler

Indiana University

OXFORD & IBH PUBLISHING CO.
Calcutta Bombay New Delhi

Indian Edition 1969 Published by arrangement with the
original publishers Houghton Mifflin
Company, Boston. U.S.A.

Price : Rs. 12 75

For sale in India, Pakistan, Burma, Thailand,
Ceylon and Indonesia.

This book has been published with the assistance of the
Joint Indian-American Textbook Programme

1

Preface

Scope of Statistics. Conventionally, the field of statistics is broken down into two divisions: (1) descriptive statistics and (2) inferential statistics. *Descriptive* statistics deals with the condensation of large masses of data into frequency distributions, measures of location, dispersion, and correlation, and other more complicated measures such as weighted indexes, standardized rates, and factor loadings. But since we often find it inconvenient or impossible to assemble completely the data which are the object of our interest and are obliged to obtain our information from a mere fragment, usually called a sample, we are compelled to infer from the fragmentary sample the probable description of the unseen population. This inductive process is not a simple task, and elaborate technique have therefore been devised to render this process dependable. These have been labeled *inferential* statistics. Inferential statistics, therefore, take their departure from the sampling of an unknown universe, which the sample will retroactively describe with varying degrees of reliability. Hence, the process and techniques of sampling beget problems of reliability, confidence, and decision-making.

Now, although the fundamental problems of descriptive statistics are only incompletely solved, still they are currently not so challenging to the inventive impulses of statisticians and mathematicians as are the problems of sampling and decision-making, which at present appear to be on the growing edge of the discipline of statistics. During recent decades, sampling and its associated issues have been the center of exploratory interest and have pre-empted the energies of creative scholars in mathematics and statistics. This concentration of interest on the fascinating frontier of statistical knowledge has been reflected in the content of some recent elementary textbooks which have even gone so far as to define the field of statistics restrictively in such phrases as "decision-making in the face of uncertainty." To such theorists, the definition of "ordering quantitative data" would perhaps seem hopelessly old fashioned.

Indian Edition 1969 Published by arrangement with the
original publishers Houghton Mifflin
Company, Boston. U.S.A.

Price : Rs. 12.75

For sale in India, Pakistan, Burma, Thailand,
Ceylon and Indonesia.

This book has been published with the assistance of the
Joint Indian-American Textbook Programme

Published by Oxford & IBH Publishing Co., Oxford Bldg.,
N-88, Connaught Circus, New Delhi-1. and printed
By R.N. Sachdev, at Allied Publishers Private Ltd.
(Printing Division) 4, Najafgarh Road,
New Delhi - 15.

common, for example, to find students embarking on partial correlations who seem unaware of the possibilities in cross-tabulation and subclassification. The exposure of inner relations does not always require such complicated manipulation; indeed, many of the complex devices conceal what relatively simpler devices may reveal.

The more leisurely and intense cultivation of fundamentals, which is one of the objectives of this text, may appear to decelerate the pace of progress of the student. In exchange, however, he will absorb a more mature and flexible understanding of statistical knowledge and a finer feeling for quantitative reasoning, as well as enhance his comprehension of sociological processes. Impatience to do "research" without such an adequate basic repertoire of tools and procedures only leads to immature applications, mechanical cookbook routines, limited and distorted comprehension of what you have when you have it, and inadequate reservations in interpretation.

Furthermore, many of the social data are of such nature and composition that they cannot profit by complex treatment and analysis, nor do they even deserve to be dressed in the trappings of recondite formulas: shabby data may take on a superficially scholarly hue and a semblance of precision which constitutes scientism of a most uaworthy sort.

Special Emphases. In conformity with the thought of assigning a somewhat higher status to descriptive procedures than seems to be current custom, we have included treatment of certain topics which are usually omitted. Among the more important and useful are standardization and norming operations. Qualitative variables, in contradistinction to quantitative variables, are given a somewhat more prominent place than is their usual lot, reflecting the currency of such data in sociology. Corresponding to this emphasis, the characteristics of attributes, their tabulation, and messures of their variation and correlation are analyzed in a somewhat extended manner.

The forms of correlation are presented in the reverse of the usual order. Instead of initiating the topic of correlation with the traditional Pearsonian product-moment method and linear prediction, the subject of relationship is launched with the simpler measures of association between dichotomous variables: Q and ϕ. The student is thereby inducted mucb more gently into the field of relationships; only wbea the concepts of joint occurrence, covariation, and marginal distribution are firmly established is the more complex Pearsonian correlation introduced. The approach to covariation is made via the principle of explained variation, which renders it at once amenable to both linear and curvilinear relations and is thus more flexible in its sociological application.

Perhaps the most significant deviation from the usual repertoire of techniques is the omission of the traditional discussion of small samples

Consistent with this conception of statistics, such authors have opened their discussion with sampling. But not being able to ignore the descriptive aspects of statistics, they *interrupt the continuity* with short digressions on the rudiments of descriptive statistics, after which they again pick up the threads of inferential statistics.

It is our contention, however, that the definition of statistics exclusively in terms of inference and decision-making confuses the part with the whole — that is, the segment which happens to be in the focus of current methodological interest with the entire field of statistical interest. What is often labeled "modern" statistics has, however, *not displaced descriptive* statistics, but merely extended it into more advanced regions. In spite of the current trends, the aim of statistics is not exhausted in sampling; the sample still serves as a vehicle for the ultimate description of the universe

It would therefore appear that the present pedagogical fashion imprudently relegates descriptive statistics to a position of secondary importance, and conveys the impression that descriptive techniques can be casually acquired. But many of the fundamentals of statistical reasoning are concentrated in descriptive procedures and are too useful in their own right to be employed merely as stepping stones to inferential procedures. Indeed, the direction of service is the reverse: the sample serves the cause of the universe. We should therefore first know what we mean by certain descriptive characteristics before making inferences about them from the sample. *It is inefficient and confusing to estimate the value of a population characteristic before the meaning of that characteristic is perfectly clear* in all of its statistical — and substantive — implications The student will simply not get the feel for the mean and the standard deviation, for example, when such concepts are accorded only incidental and passing treatment.

In this text, therefore, descriptive statistics are first presented separately, *uncomplicated by reference* to sampling procedures, which are sufficiently difficult to acquire even after descriptive statistics have become second nature. We think that the beginning student will appreciate such consideration.

The premature emphasis on inferential, to the neglect of descriptive, statistics has the further unfortunate side effect of an overhasty adoption of complex measures and formulas which are only vaguely visualized, and certainly not adequately *appreciated* *There seems* to prevail among many students, who have acquired some facility in the middle range of statistical training, a tendency to use their most "advanced" knowledge, when more thorough probing of simpler devices would be more effective *and meaningful*. This inclination to embrace complicated models, which are often much less effective than more elementary methods in displaying the data, actually inhibits sound quantitative reasoning. It is not unmachines, without the intimate familiarity in miniature that is acquired by hand calculation, would lead only to rontinistic, mechanical, and robot statistics.

Machine calculation, it is therefore contended, has little place in the elementary course which emphasizes fundamentals. Not only are the problems not so elaborate as to require a machine, but the entire emphasis should be on the acquisition of principles and the encouragement of quantitative reasoning, uncluttered by excessive arithmetic calculations which add nothing to statistical theory. In fact, complicated machine calculations will only becloud the internal issues for the immature worker. Computing outfits are no more meticulous than sausage machines in clarifying the materials that pass through them, unless a knowledgeable person selects the ingredients, orders them, and creates proper mixtures.

Mathematical Prerequisites. This text joins the increasing number of those that do not make demands beyond the ordinary high school mathematical background. Indeed, such a reassurance has become so standard in the textbook blurbs that it has almost attained the status of a cliché. The apologetic reason often given for this academic practice is simply that the college student does not possess mathematical competence beyond high school algebra anyway, and we are therefore forced to adjust ourselves to the deplorable reality.

It is unquestionably true that students are able to enter the university with inadequate competence even in arithmetic, to say nothing of algebra and trigonometry. However, it is not a question of surrendering to the laxities of high school education, for there are more constructive reasons for the alleged disregard of mathematical requirements in elementary statistics. It seems to the present authors, as well as to others, that much statistical reasoning may actually be competently carried on with only a good arithmetic background. Furthermore, many mathematical formulas may be put to adequate and profitable use without a comprehension of the technical means by which these formulas are derived. Consequently, if the formulas are not to be derived anyway, higher mathematical training is less essential to their implementation. In fact, a considerable degree of statistical skill can be developed on, at most, a base of elementary algebra, at least during the first — and possibly second — year of academic training. Such a concession does not argue affiliation with the "statistics-made-easy" school of thought. Formal skills must be mastered if they are to be imaginatively applied to social data. On the other hand, the mechanically flawless application of such procedures without regard for their relevance to subject matter can readily lead to nonsensical and deceptive conclusions. It is the aim of this book to cultivate the art of statistical reasoning by focusing on the union of neutral arithmetic routines and concrete subject matter.

and "Student's" *t*. This was deliberate, although such a decision is recognized as harboring some reservation. Small-sample theory may be viewed as an amendment to the techniques of sampling procedure rather than as a new body of theory; and it has been the principle of selection for this text not to enter into the manifold variants of standard procedures, but rather to concentrate on the thorough understanding of fundamentals within the scope of an elementary framework.

Mathematical notation has been employed as sparingly as possible. Instructors are not always fully aware of the frustration experienced by the students in acquiring a new "alphabet." Symbols are never introduced in anticipation of their later use, but only when actually required.

By far, most of the data employed in illustrating statistical processes are taken from sociological and related sources. It is, of course, obvious that such meaningful materials elicit and reinforce student interest, which might otherwise tend to lag. However, in certain instances, hypothetical data of a very simplified sort have been introduced for experimental demonstrations of the behavior of data because they are so easily maneuvered, and more clearly expose the procedures under examination. Such imaginary data actually become more meaningful than the collected empirical materials, which are often too cumbersome for the purpose.

In statistics, as in other disciplines, a certain amount of pedagogical energy is devoted to mere terminological problems. This is inevitable, since the only channel of communication is terminology. Not until an author has examined several hundred standard references, in the process of increasing that number by one, does he realize the variation in the use of concepts. Fortunately, much of the controversy on terminology is merely terminological and does not alter one whit the statistical practice. We have attempted to codify what seems to be good practice, but have veered therefrom at times for purposes of consistency. Occasionally such deviations have been briefly noted.

Machine Calculation. An impression prevails in some quarters that desk calculators and electronic computers are tending to displace pencil-and-paper statistics, as cultivated in such a text as the one before us. No greater self-deception is conceivable than the one that machines have rendered obsolete the old-fashioned virtues of hand calculation. The typewriter and printing press have not driven out of existence the slower methods of penmanship; pencil and paper have not become vestigial tools. There are several reasons for the insistence on pencil-and-paper methods in this text. Many students, as well as many workers in the field, do not have electronic computers available, or even desk calculators. Nor are most of the problems encountered in ordinary affairs sufficiently elaborate to require such expensive apparatus. Even if electronic machines were the ultimate answer to all problems, it would still be true that the use of

Acknowledgments. While every book is truly n group effort, it is also true that some individuals contribute more tangibly than others. Foremost among the contributors to this book are the many students whose thoughtful reactions enabled us to see more clearly what issues are most troublesome to the beginner. Also high in our debt are those graduate assistants in social statistics whose critical examination of study problems led in many instances to revisions of those expository portions of the text which appeared to be wanting in clarity or completeness. In particular, it is a pleasure to thank Cherry Carter, Herbert Costner, and Donald Horning, who attended to such a large part of the numerical detail that necessarily enters into a statistical work hut which is often inconspicuous in the final product. We are especially indehted to Professor Peter Rossi, whose critical but sympathetic reading of the entire manuscript issued in numerous constructive suggestions.

We are indebted to Professor Sir Ronald A. Fisber, F.R.S., Cambridge, and to Dr. Frank Yates, F.R.S., Rothamsted, also to Messrs. Oliver & Boyd Ltd., Edinburgh, for permission to reprint Table IV from their book *Statistical Tables for Biological, Agricultural and Medical Research.* Finally, to the following publishers we are ohliged for their permission to reproduce copyrighted materials: *The American Journal of Sociology, The American Sociological Review,* Chapman and Hall Ltd., Columbia University Press, *Educational and Psychological Measurement,* Harper & Brothers, McGraw-Hill Book Co., Inc., Metropolitan Life Insurance Co., Prentice-Hall, Inc., Princeton University Press, Simon and Schuster, Transactions of the Society of Actuaries, the University of Chicago Press, John Wiley & Sons, Inc., and Yale University Press.

JOHN H. MUELLER
KARL F. SCHUESSLER

Bloomington, Indiana
November, 1960

Sociology and Statistics. This text has been prepared primarily for sociologists, since other disciplines will necessarily select different elements. Furthermore, sociology has lagged behind economics, psychology, and education in its cultivation of statistics. This unfortunate lag expresses itself in the more or less independent development of statistics outside the scope of the basic courses in the department usually offered on the lower levels and even in graduate divisions. Unlike economics and psychology, the introductory courses in sociology assume very little in the way of *statistical knowledge or training.* Indeed, in many instances, the first elements of statistics are not even embarked upon until the student has attained graduate status. This text seeks to break down that compartmentalization by gearing the presentation to the undergraduate, with courses on middle levels primarily in mind; but it is hoped that it will contribute to the cultivation of quantitative methods among students of all ranks, including, of course, the graduate student without previous statistical training.

It is recognized that many students in sociology are not destined to become professional workers in the field. By far, most of them will become consumers, and not producers, of statistical materials. We therefore believe that the mastery of introductory statistical techniques and the comprehension of the principles involved will carry with it cultural benefits similar to those gained in the pursuit of substantive courses.

And even professionally, many graduates will discover that descriptive statistics are sufficient for their purposes and that inferential statistics are only rarely called into play. Their primary need will therefore be for those skills which enable them to "move" their field observations onto the printed page by compiling them and presenting them in a manner appropriate to their purposes and faithful to the nature of the events themselves — in short, in such a way that the statistics do not "lie." Hence, the beginning student should not be deluged by a confusing number of variants of the same procedure, but should be inducted into a limited number of fundamental methods, which should be accompanied by rational analysis to ensure competent application to social data. Consistent with this point of view, the references at the close of each chapter are appended, not as a pool of highly technical information, but rather as a source of greater enlightenment on the role of statistics in social life and social science.

This book had its inception and development in the experience of teaching statistics to undergraduate and graduate students in sociology and the allied social sciences over a period of years. Many of the questions treated in this text in a somewhat expansive manner took their origin in the queries of the more reflective students. The teacher will, of course, proceed at his own pace, mindful of the stated objectives of the text and also of the ability of his students.

Contents

We are pleased to acknowledge the help of those readers who called our attention to typographical and computational errors in the first printing of this book. Although such mistakes were not numerous, we have taken advantage of this second printing to correct them, and we are confident that students will accordingly be less distracted by minor flaws.

John H. Mueller
Karl F. Schuessler

Statistical Reasoning in Sociology

Statistics in Social Life 1

The Individual and the Collective. If social events, such as marriages, births and deaths, crimes and delinquencies, or public opinions were no more numerous than chairs in the living room or children in the family, there would be no great difficulty in apprehending them. In fact, there would be relatively little need to generalize about them, for the human intellect could encompass them separately and individually. But when, instead of a small number of easily identifiable objects, we are called upon to treat large aggregates, individualization becomes mentally impossible. Consequently we can view them only collectively, in terms of one or more common characteristics, and with corresponding incompleteness. The shepherd of Biblical antiquity "who calleth his own sheep by name" must have had small flocks indeed. Today, on the Australian or American ranch, the large flock would be counted, classified, and summarized in various ways, in the course of which the individuals would lose their identity in the total anonymous mass. So inadequate has the individualistic approach become, that, for purposes of description or prediction, various devices of summarization have been evolved; or a limited number of individuals (a sample) are selected from the aggregate to represent the total, about which generalizations may be made with varying degrees of precision. Such economical numerical procedures have, of course, been in use for centuries; but during the last two hundred years, these procedures have become formalized and have come to be known as "statistics."

This evolution in the history of culture from naming to counting was neither a product of arbitrary and spontaneous personal choice nor a private accomplishment. It was rather a matter of gradual compulsive social adaptation to the exigencies of an expanding social organization which had become increasingly immense and complex, in which symbols in the mid-twentieth century by the National Safety Council. Quételet feelingly comments on his epoch-making discoveries:

> Thus we pass from one year to another . . . seeing the same crimes reproduced in the same order, and calling down the same punishments. . . . We might enumerate in advance . . . how many will be forgers, how many will be prisoners; almost we can enumerate in advance the births and deaths that should occur.*

In mid-century, Henry Thomas Buckle (1821–1862), the English social historian, caused similar consternation and moral indignation among his incredulous pious readers by calling attention to the fact that, year after year, about 250 persons committed suicide. After pointing out that, among public crimes, none seems to be so completely dependent on individual impulse as suicide, he still observed that

> it is surely an astonishing fact that all the evidence we possess leaves no doubt . . . that suicides are merely the product of a general condition of society, and that the individual felon carries into effect what is a necessary consequence of preceding circumstances. In a given state of society a certain number of persons must put an end to their own life.†

There has also been a secular version to these objections. Although such opposition has by now been almost dissipated, it was once thought that social science should not hope to emulate the mechanical methods of the physical sciences by treating social behavior quantitatively. It should rather restrict itself to the operations of mental phenomena such as understanding, insight, empathy, sympathetic introspection, and other subjective techniques that the German theorists have subsumed under *Geisteswissenschaft.* Such intellectual methods, it is averred, are appropriate to the true nature of society, which has its being in subjective communication Statistics, it is still thought, somehow debases our observations and dehumanizes society. It is an unnatural imposition of external techniques upon social realities, which exist in the cultural imagination and are therefore essentially unquantifiable. Statistics squeezes human behavior into alien categories, and is an illegitimate rival to other methods which recognize the true nature of humanity. Strangely enough, many persons who casually assert that "most people" "usually" behave in such and such a manner, and who predict that the "chances are" so and so, are still offended at the thought of pinning down human behavior to quantitative terms.

Further "dangers and fallacies" of statistical method have been belabored by its critics. It is said that statistics can offer only probabilities,

* F. H. Hankins, *Adolphe Quételet as Statistician*, New York, 1908, p. 83.

† John Venn, *Logic of Chance*, London, 1876, p. 467.

became a substitute for the individual object itself. Statistics, therefore, is an instrument for viewing the mass of happenings around us. To be sure, in our private life we are still concerned with personal opinions, marriages, deaths, and many individual transactions. But in our social, collective relations, we are concerned with public opinion polls, population and vital statistics, actuarial descriptions, and predictions, on the basis of which society pursues its collective interests.

To accomplish these ends, it has been necessary to develop a number of quantitative techniques which serve not only to describe coherently the collectivity of life around us, but also to anticipate, or forecast, their recurrences in the future, which we express in terms of degrees of probability

The concept of statistics may therefore be defined in two related senses: (1) the factual data themselves, such as vital statistics, statistics on trade, production, and the like; and (2) the methods, theories, and techniques by means of which the collected descriptions are summarized and interpreted Although the procedures of statistics, as they have been elaborated and formalized, may sometimes appear to the uninitiated as recondite, and remote from social interests, they are in fact merely an extension of what every intelligent person does anyway. They have become a personal and social necessity, reflecting our need for understanding the past and anticipating the future, as they involve the treatment of large masses of phenomena. Hence, the field of statistics has sometimes been characterized as the "science of large numbers," in which large volumes of data are systematically compressed into smaller and more manageable compass

Validity of the Statistical Method. The statistical method is not only a set of practicable techniques, but it is also an ideology which validates the use and application of these techniques. There are, of course, other views of nature besides the statistical, as is evidenced by the opposition which this method has aroused in some quarters during its history, especially when applied to human behavior. By some moralists, for example, the endeavor to reduce human behavior to statistical regularities has been condemned as materialistic and as a denial of the axiomatic free will of man To set forth the quantitative regularities of inanimate nature is one thing, but to reduce the behavior of men's souls to mechanical laws is to undermine the very foundation of personal responsibility and morality. Quételet (1796–1874), the Belgian statistician, was among the first to apply these quantitative procedures to human behavior in citing the constancy in the number of crimes from year to year. This regularity, he asserted, could be used as a basis of probabilistic prediction — in the commonplace manner that automobile deaths are forecast for holiday weekends

In the modern sense, statistics began to be linked with mathematics and to take on its early characteristics in the seventeenth century with Blaise Pascal and Pierre de Fermat; but it assumed its more mature physiognomy in the eighteenth and nineteenth centuries. Abraham de Moivre discovered the normal curve around 1730, about the time of Newton's death (1727). Although descriptive statistics, in a kind of primitive way, dates back to the census of the Biblical epochs — a fact which hardly detracts from its contemporary importance — probability statistics had to await the emergence of other social needs and interests.

The laws of probability, which together with descriptive statistics make up the basic dichotomy of current statistical method, had their beginnings in two quite different social circumstances: (1) the study of the problems of gambling, which was a pastime of the French nobility; and (2) the actuarial interests of the British commercial bourgeoisie. With some oversimplification in statement, the former gave rise to what is now known as classical *a priori* probabilities; the latter to empirical probabilities. With the apparent intention of eliminating some of the risk from gambling, the eighteenth century gamblers consulted mathematicians like Pascal, de Moivre, and Daniel Bernoulli to illuminate the outcomes of games of chance. On the other hand, among the merchants of Puritan England, the motivation for statistical enlightenment was perhaps on a more respectable plane. Interested, as they were, in the growth of capitalism and in demography, they cultivated problems of an actuarial nature, which became the basis of their statistical calculations. As early as the seventeenth century, the Englishman John Graunt (1620–74) was the first to measure the regularities in the duration of human life, and to consider the relation of birth and death rates to occupations and business conditions. He found fairly constant demographic rates which were among the first secular evidences of an orderly plan in nature for population behavior. In 1693 Edmund Halley published the first mortality tables. By the beginning of the nineteenth century and beyond, Pierre de Laplace in France, Adolphe Quételet in Belgium, Karl Friedrich Gauss in Germany, and Sir Francis Galton and Karl Pearson in England, were adding important increments to the science of probabilities from both the mathematical and empirical sides.

Thus, the practice of statistics arose out of the convergence of two sources: (1) descriptive data of a political, economic, and demographic *character, and (2) the mathematical researches in probabilities. These* were supplemented toward the end of the eighteenth century by astronomers, who developed the theory of errors out of their astronomical measurements. Quételet was the first to fuse these interests for social purposes. In the latter part of the nineteenth century, biology and genetics began to contribute significantly through the work of Francis Galton and Karl Pearson and R. A. Fisher, who are now known for their development of not certainty; that they apply only to the mass, not to the individual. The answer is that these assertions are unreservedly true. But they should be viewed as characteristics of the method, rather than limitations. An invidious distinction has thereby been drawn between statistical forms of knowledge, and other forms of knowledge, as though these other forms of human knowledge did not also fall short of omniscience.

For a person who has profited by statistics, it is difficult to share such skepticism. The statistical method, rather than replacing "insight," reinforces insight. No person can become less human by employing statistics, nor more human by avoiding them.

As a matter of fact, there is a continuity between common sense, which informally makes rough quantitative judgments, and statistics, which is not only a more formal and precise version of such knowledge, but is also of more extended scope. Although many a person may "lie with statistics," other persons who eschew statistics probably lie just as effectively without them. The "tyranny of numbers" is probably no more mischievous than is the "tyranny of words." The subjective and the quantitative approaches have been learning to coexist and to supplement each other.

Evolution of Statistics. The most prevalent early consumers of mass statistics were the kings and princes, whose vital concern lay in the wealth of their domain: the potential soldier population, its agriculture and manufacture, the financial resources to support their military and civic enterprises; hence, the etymology of the term in the concept of "state." However, the present term predated its quantitative connotations. In Elizabethan England, the "statist" was a political functionary or statesman, whose repute apparently corresponded to the modern term "politician"; Hamlet avers that

> I once did hold it, as our statists do,
> A baseness to write fair . . .

In mid-eighteenth century Germany, *Statistik* was the study of state, or public, administration, as distinguished from the history and philosophy of the state. This concept was later carried to England as "statistics," and in the early nineteenth century to the United States. Thus "state arithmetic" — sometimes called "political arithmetic" in the seventeenth century — evolved into "statistics" in a perfectly normal linguistic manner, destined to rhyme with ballistics, logistics, and other concepts denoting a systematic analysis of areas of practical human knowledge. Although governments are still voracious consumers as well as energetic producers of statistics, the concept has long ago been extended to innumerable other areas of activity in the physical and social world and has lost almost entirely the earmarks of its political origins.

abstract, formula. These will never quite conform to each other, and sometimes, in fact, are only a rough fit — and this for various specific reasons. (1) Often the data are only crudely quantifiable, as, for example, attitudes and opinions. (2) Every summarization of data can be only a compilation of one, or a few, aspects of the observed objects; these aspects are, so to speak, torn out of context. But the factors that are excluded from the formula do not thereby cease operating! (3) Formulas vary in their sensitivity to the complexity of the data and pick up only limited aspects of their manifold characteristics.

Thus, the data must be prepared and groomed for the formula, which must be fed the data in a form that it can digest. For example, a formula cannot absorb subjective attitudes, but will accept the data only when in the form of measured behavioral equivalents. A formula is very like a machine, which should not be asked to do a piece of work for which it is not adapted; otherwise, the formula will chew up the data which are dumped into it and produce a chaotic mass of uninterpretable digits. Hence, the application of statistics requires continued discretion grounded in the knowledge of the subject matter to be interpreted.

Such emphasis on quantitative reasoning will excite in the student a curiosity about the solutions to the realistic problems and thereby enlarge his sociological imagination. For it can hardly be expected that every student in the social sciences would develop a "mathematician's delight" in the contemplation of the elegance of an abstract, mathematical procedure, any more than a mathematician will work up excitement in unraveling divorce and delinquency rates.

Civilized man is literally engulfed with statistics: public opinion polls, cost of living indexes, population data and vital statistics, actuarial calculations in life insurance, games of chance, blindfold tests, and many other evidences of the prevailing quantitative view of life. The common man, in his unsophisticated way, employs crude statistical concepts when he speaks of averages, hunches and hypotheses, probabilities, chance, long run, samples, and of "all other things being equal." Statistics, in the sense that it involves skill in thinking, is therefore not a redundant academic accomplishment, but is an integral element in the current thoughtways of our civilization. The development of the contents of this volume and the selection of the problems and questions for thought have been designed to foster those skills in statistical reasoning which are essential not only to the professional social scientist but to the intelligent citizen as well.

Questions and Problems

1. Consult the dictionary and other reference books for definitions of statistics.
2. Distinguish between numerology and quantification; quantification and statistics.

theories of association and correlation and other techniques of measurement

In retrospect, it is profitable to contemplate the diverse seminal sources of statistical methods. They received their impetus from a wide variety of social practices and intellectual disciplines: demography, commerce, games of chance, astronomy, biology, agriculture, and mathematics. They have now extended their applications to linguistics, anthropology, communications theory, aesthetics, and other disciplines. They either touch lightly, or are heavily involved in, almost all of the conventional scholarly disciplines

Principle of Statistical Reasoning. Statistics, when competently cultivated by the social scientist, comprises much more than the manipulation of figures and formulas. Statistical procedures, when applied, consist in relating or fitting social data to the appropriate statistical formulas and equations. But even this can degenerate into (or not rise above) the handbook approach of merely plugging the data into some formula and routinely solving the equation. Effective application consists in the careful substantive analysis of social data for their adaptability to the statistical techniques, as well as the careful assessment of the technique for its potentialities in processing the data, to the end that the model and the data may be mutually compatible and neatly joined. The two frames of reference are coordinate and inseparable; hence, in order to implement the application, familiarity with both the statistical principles and the subject matter are equally essential. In this text, considerable emphasis has been laid on the fusion of elementary statistical procedures with the social data; for this collaboration between the two disciplines constitutes the very basis of what we shall call *statistical reasoning.*

Such reasoning is made necessary by the fact that a statistical formula or model never quite fits the social data; nor do the empirical social data necessarily fit the ideal procrustean structure of the statistical models. We should not be duped by the external precision of a pat formula or by the "exactitude" of numbers. A dozen eggs do not fit the uniformity of the arithmetic digits which represent them; the distribution of a finite number of heights or weights does not quite attain the smoothness of the abstract normal curve which assumes an infinite number; the rank-ordering of choices does not result in equidistant ranks, as the numerical ratings may imply. Whatever the readjustments or rearrangements we may undertake for greater precision, the abstraction still remains an abstraction that is not a duplicate of the observed empirical objects. It therefore lies within the very nature of the statistical method and the nature of reality that the two do not exactly coincide.

Thus, the solution of any statistical problem involves the dovetailing of more or less imperfectly observed fragments of data with an ideal,

Social Variables and Their Measurement 2

Section One

Fundamentals of Measurement

Social Need of Quantification. Any event or object in nature may be considered as being made up of a number of component factors: the chemical contents of the human body, the ingredients of a cake, the elements in the production of a disease, the social factors that produce a war, or the dimensions of a social institution. It is the aim of the physical and behavioral sciences to analyze an event in order to make it understandable. However, it is not enough merely to identify the factors which enter into an event; it is also necessary to measure their force, intensity, or quantity. A chemical formula, a blood count, the amount of a medicinal dose, the intensity of a social attitude, the size of a population — all testify to the principle that measurement is essential to knowledge. *Without measurement there would be no statistics.*

But measurement implies that we have units of measure. How to devise effective units of measure, and how to make nature stand still long enough to apply these devices, are basic problems in both the physical and social sciences. The typical laboratory method, by which nature can be placed under a microscope or poured into a beaker and heated to the boiling point, to say nothing of the utilization of enormously more complicated apparatus, is an invention of the physical scientist for the express purpose of immobilizing nature long enough to be manipulated, observed, and measured.

It is frequently asserted, and perhaps with some support, that social data are collected and quantified with considerably more difficulty than are physical data. There are, indeed, many obstacles interposed between the social scientist and his ultimate measures: the data often cover a large

3. Elaborate the statements: "Statistics is an extension of common sense"; "Every man is an informal statistician."

4. Compare "the statistical view of the universe" with other conceptions of the structure of the universe.

5. List and discuss social factors in the growth of statistics in Western civilization.

6. Formulate the argument that statistical methods cannot provide valid knowledge about man and society.

SELECTED REFERENCES

Cohen, Morris R , *A Preface to Logic* Henry Holt and Company, New York, 1944. Chapter 7

Huff, Darrell, *How to Lie with Statistics.* W. W. Norton & Company, Inc , New York, 1954

Kendall, Maurice G., and William R. Buckland, *A Dictionary of Statistical Terms* Oliver & Boyd, London, 1957.

Kline, Morris, *Mathematics in Western Culture* Oxford University Press, New York, 1953. Chapters 22–24.

Lundberg, George A., "Statistics in Modern Social Thought," in *Contemporary Social Theory* Edited by Harry Elmer Barnes, Howard Becker, and Frances Bennett Becker. D. Appleton-Century Company, New York, 1940. Pages 110–139.

Mahalanobis, P. C , *The University Teaching of Social Sciences: Statistics.* UNESCO, 1957.

Walker, Helen, *Studies in the History of Statistical Methods.* The Williams & Wilkins Company, Baltimore, 1929.

Wallis, W Allen, and Harry V. Roberts, *Statistics· A New Approach.* The Free Press, Glencoe, Ill , 1956

Willcox, Walter F , "History"; Robert M. Woodbury, "Statistical Practice"; Oskar N Anderson, "Statistical Method"; under "Statistics" in the *Encyclopædia of the Social Sciences.* The Macmillan Company, New York, 1934. Volume 14

and female; a person's marital status may be single, married, widowed, or divorced; nationality may be American, French, or Italian. The attributes of a given qualitative variable cannot be scaled, or arranged in order of magnitude. A given sex, nationality, or marital status cannot be considered as being "higher," "greater," or "larger" than another in their respective series.*

Because of this limitation, it might appear that the statistical method would have little to contribute to the analysis of qualitative data. But quite to the contrary: there is a large and growing body of statistical techniques now available for the treatment and manipulation of qualitative variables. These are of particular importance in sociology, simply because many significant sociological data are qualitative in form.

The corresponding terms in the above classification may be succinctly set forth as follows:

Variables	*Values*
Quantitative	Variates
Qualitative	Attributes

Nevertheless, in its application, even a simple classification of variables as quantitative and qualitative will turn up apparent ambiguities, because some traits may seem to fit both types. Thus, occupation, religion, and crime are undoubtedly qualitative variables which are not intrinsically measurable in units as are age, weight, or size of population. And yet we could consider religious denominations as more or less orthodox; crimes may be ranked in order of seriousness and by severity of punishment; and occupations may be scaled according to their social prestige. Such measures, however, are not intrinsic to the attributes; they merely reflect the separable social values that often adhere to such categories. Different states classify crimes differently; communists and capitalists would rank occupations in different order.

In reverse fashion, we often attach what appears to be a qualitative concept to clearly quantitative data. Various age groups may be quali-

* Some authors apply the concept "variation" only to variables that can vary in quantity. Qualitative data are therefore sometimes labelled as "non-variable" or "enumerative," emphasizing thereby that qualitative categories cannot be measured as magnitudes, but only enumerated as qualitative traits. Magnitude and quality would therefore be antithetical concepts. This text, however, shares the position of those who define variation as covering both quantitative and qualitative characteristics.

This disputation in terminology can hardly be dignified as a semantic problem, since there is never a question of essential meaning if terms are consistently employed. It is simply a problem in labeling. The nomenclature employed by statisticians is by no means uniform, not only as applied to concepts here under discussion, but to other statistical concepts as well. This circumstance leads to confusion in the minds of students who obtain their statistical education from a variety of academic sources In subsequent instances, this lack of uniformity will not always call for comment, but the student should constantly be alert to that possibility.

social area; their collection may violate the sense of privacy of the subjects; they may be subjective in nature and hence difficult to objectify. Furthermore, society is heterogeneous and dynamic, and many of its dimensions do not lend themselves too well to standardized units or categories. Thus, social distance is more difficult to measure than geographical distance; social forces are more elusive than are physical forces.

Nevertheless, we must start with the basic assumption that anything that exists must exist in amounts of more or less, or with *greater* or *lesser* frequency. Accordingly, there are innumerable legitimate questions which social scientists are called upon to illuminate How, for example, can one count the number of Negroes in the United States when their numbers are changing every day, and the very concept of race is subject to a variety of definitions? How may we secure measurable data on the simple assertions that Americans are marrying younger now than several decades previously; that there is less unemployment now than in the past; that crime increases in postwar years; that race prejudice is decreasing; that men are better drivers than women, or that the taste for modern music is rising? How are we to count and measure the factors in juvenile delinquency or in divorce?

These questions have these features in common: (1) they call for objective, quantitative answers, and (2) the appropriate units of measure are not immediately apparent and are almost always difficult to devise. Broadly speaking, the manner of quantification is governed by the nature of the variables under consideration.

Quantitative and Qualitative Variables. Any object or event which can vary in successive observations either in quantity or quality may be termed a *variable*. Variables are accordingly classified as quantitative or qualitative.

A *quantitative* variable is one which may take on various magnitudes, i e, may exist in greater or smaller amounts. Examples of quantitative variables are. age, height, income, size of population, size of family, length of prison term, birth rate, and numerous other characteristics. All persons possess the trait of age, but some have a higher and some a lower age; cities are of varying sizes; convicts serve prison terms of varying lengths. These variables can be measured, and each resulting magnitude is called a *variate* A set of variates, such as the heights of the individual soldiers in a platoon, or the respective sizes of 1,000 families, can be arranged in order of magnitude from the smallest to the largest. We can locate them on a scale. This inevitable order of the data can be accepted as a convenient earmark by which a quantitative variable can be distinguished from the qualitative type.

A *qualitative* variable may also vary in successive observations, not in magnitude, but rather in quality or kind; such qualities are customarily called *attributes*. Thus, sex will vary according to the attributes of male tatively designated, in order, as "infants," "children," "adolescents," and "adults"; letter grades may be substituted for raw test-scores; the United States Census calls places with fewer than 2,500 inhabitants "rural" and all larger places "urban." But no one should be misled by these labels into believing that the data have thereby been transmuted from quantitative into qualitative; we have merely attached new verbal symbols to arbitrary segments of the data. Anyone familiar with the symbolism will still interpret the data quantitatively. These labels, which merely stand for intervals of varying width, are pragmatically more convenient than the precise raw measures. The mere reduction in precision does not change the nature of the data.

Discrete and Continuous Variables. Statistical procedures differ not only according to whether the data are qualitative or quantitative, but also according to whether they are discrete or continuous. *Discrete* variables are based upon events which are considered indivisible, which do not vary in amount, and which are therefore merely present or absent. We cannot fractionate attributes: a person does not commit half a crime, or secure a fraction of a divorce, or hold one-third of an occupation. An event is either a crime or not a crime, either a divorce or not a divorce. Since there are no gradations, attributes have no natural sequence or continuity; there is no scale distance between them; it is only possible to "jump" from one value to another. It is obvious that all qualitative variables are by nature discrete.

Quantitative data, on the other hand, may be either discrete (discontinuous) or continuous, which explains why the distinction is usually reserved for them. In this restricted sense, a discrete variable always involves *counting* the number of events, for example: the number of persons in each family, the number of whole days each worker is absent, the number of inhabitants in each city, and so on. As these examples suggest, a discrete variable consists only of whole numbers, and fractional values cannot occur.

Continuous variables, on the other hand, are theoretically infinitely divisible into smaller and smaller fractional units. Age, distance, weight, intelligence quotients, and various kinds of rates can take any one of the innumerable small values on the scale, or continuum. In order for a person to pass from eight to nine years of age, he necessarily passes through every minute gradation between the respective years and at some time has occupied every one of the values on that continuum, even though he cannot pause to measure them.

Hence, a measure of a continuous variable can never be considered exact. However accurate the measure, we can always conceive of the hypothetical possibility of a still more precise measure. The one-thousandth of an inch on the micrometer, the split second in a photo-finish of described. Hence, we must necessarily rely on measures of things which *are* accessible and observable, and which may be accepted as *operationally equivalent* to tbe subjective variahle which is inaccessible to direct measurement.

To the question, "Is racial antipathy toward tbe Negro declining?", we would in effect reply: "We cannot penetrate tbe psyche of tbe population directly, we are not 'mind readers'; but we can observe and measure the annual lynchings, the rate of desegregation, tbe occupational changes in the Negro population, the content of editorial expression in journals and newspapers, and other behavioristic manifestations which must be presumed to be the equivalent of the subjective attitudes of tbe population in question." Similarly, elusive phenomena, such as bappiness in marriage, have been measured by such behavioral analogues as duration of marriage, the collaborative activities of the spouses, the number of tensions incurred. Or, again, changes in musical taste are measured hy tbe content analysis of the concert repertoire to which the audience attends, or the volume of specified records which it purchases.

This indirect procedure in measurement is by no means restricted to social science. In physics, heat is measured indirectly hy.the expansion of a column of mercury; electricity is also measured not in terms of itself hut by the work it does. We can measure only what can be perceived hy the senses.

Research scholars in the social sciences have exercised considerable ingenuity in devising equivalents for the measurement of the intangihles of social life. But such circuitous measures are encumhered with certain prohlems from which direct measures are quite free. It is important to realize that equivalents are not identities. Among the many alternatives which could he proposed, no single one, nor several in combination, will be a perfect replica of the thing they presumably represent. In the study of racial antipathy, the number of lynchings, interracial club memberships, intermarriage, and otber correlates are not equally sensitive in registering tbe degree of positive or negative attitudes.

In spite of tbese and other difficulties, there is nothing essentially new or esoteric in the employment of behavioral equivalents. Common sense bas always made inferences about individual motives, attitudes, and values on the basis of observable behavior. In fact, the aphorism that "actions speak louder than words" is a recognition of the view that actions may be more reliable as attitude indicators than mere "words," which among honest people presumably reflect tbeir basic attitudes and beliefs.

(4) Lastly, measurement may consist in *ranking* a series of objects according to a selected criterion. Such *ordinal measures* may be constructed on the basis of the objective measures already described; or they may be purely subjective in character, with little or no overt indication of the basis of judgment. Thus, in a relatively objective manner, cities

Although these issues may seem somewhat trickily pedantic to some readers, the concepts of continuity and discontinuity are basic to numerous statistical interpretations.

Devices for the Measurement of Observed Data. From the fact that variables are not uniformly constituted we have already been able to discern that no single method of measurement will work on all types of data. We here formulate four measuring devices, presented in order of objectivity and amenability to replication: (1) direct enumeration; (2) application of abstract standard units of measurement; (3) behavioral equivalents of the phenomena under observation; and (4) ordinal ranking.

(1) *A direct enumeration is a simple count of items* in terms of themselves. The categorical item itself is the unit of measure. Thus, the number of divorces, crimes, persons, or the number of rights and wrongs in a questionnaire are all direct enumerations of the defined units. All of the counted items are considered identical, while their differences and variations are ignored In fact, in order to be additive, the items *must* be viewed as identical. This is the simplest conceivable type of quantification, familiar to every boy who counts his marbles.

(2) As already intimated, a direct count sometimes lacks desirable precision because it ignores the variation between items. Bananas, which vary in size, may be sold by the dozen (a direct enumeration), or by the pound. By converting bananas to pounds, we employ, on a slightly higher level of abstraction, a *standard unit of measure* against which the objects are matched and compared. The pound, the inch, the year, are conventional norms of measurement understood by all members of our society. They have replaced the more primitive units of comparison, such as the hand, the foot, the thumb, the cubit (forearm), the pace, and horsepower — some of which still linger in folklore. But none satisfy the requirements of precision in modern commercial transactions and technology.

Both of these types answer to the requirement of objectivity and are readily subject to replication by other observers. They produce what some have named *ratio scales*, since relative intervals that are equal may be expressed by the same ratio.

(3) The third measuring device utilizes what we shall call the *behavioral equivalent*, which yields measures that are less direct and less valid than those of the forenamed procedures. In the social sciences, much of the subject matter consists of patterns of attitudes and values which are significant for the understanding of human behavior. However, these intangibles are not accessible to material frequency counts or palpable measuring sticks Public opinion, social attitudes, social status, marital happiness, race antipathy, group morale — all constitute legitimate subject matter for sociological investigation; but their subjective, cultural character does not permit the application of the direct measuring devices hitherto

Attribute
Discrete Variable
Continuous Variable
Direct Enumeration
Standard Unit of Measure
Behavioral Equivalent
Ordinal Ranking

2. Family size is a discrete variable, yet the average U.S. family has 3.7 persons. Reconcile these two statements.

3. Family income is measured in discrete dollars and cents, yet we have defined income as a continuous variable. Explain this seeming inconsistency.

4. Indicate whether the following variables are discrete or continuous:
   Attitude toward war
   Size of city
   Suicide rate
   Slot-machine payoff
   Age
   Number of births each year

5. Give several examples of discrete and continuous variables from the field of social science not mentioned in the text.

6. List several social variables that cannot be measured directly, but which could be measured by a behavioral equivalent.

7. Individual A endorsed 50 out of 100 items reflecting racial prejudice, while B endorsed 25. Is it reasonable to conclude that A is twice as prejudiced as B? Explain your answer.

8. A group of students were asked to rank three professors according to teaching ability. What problems are involved in this measurement technique?

Section Two

Errors in Measurement

Numbers and Things. The classification of social variables and the available devices for their measurement have been portrayed, not as inherent categories of nature, but as cultural inventions for the purpose of observing and ordering the world around us. Since measurement is a human activity, all measures of such variables are necessarily infected with human fallibility. If the methods of measurement in current use constitute improvements over those of a century ago, the various alternatives now available will in turn be replaced by modifications yet to be devised.

To be sure, mathematical quantities, in which all measures are couched, have the external appearance of being exact and absolute — a fact which may simply be arrayed in order of size after a careful count of their inhabitants, or married couples may be ranked for happiness after elaborate marriage-adjustment scores have been computed. In such instances, mere ranking involves throwing away carefully calculated accuracy, which may not be needed at the moment. On the other hand, in a predominantly subjective manner, pupils may be ranked by their fellows in order of personal preference, or pictures may be ordered by judges who doubtless apply diverse criteria, even though these criteria remain more or less unconscious, or at least unexpressed Analogously, occupations, neighborhoods, poets and musicians, may be ranged in order of prestige. In general, any series of *attributes* may be ranked according to some external criterion; and *variates* arranged according to their respective magnitudes.

It must be obvious that the statistical difficulties in ranking lie in its relative coarseness Owing to its judgmental nature, (1) intervals between successive ranks are not necessarily identical, and (2) two or more independent rankings of the same objects will not necessarily be comparable. However, when intuitive norms are the only ones available, or when precision is of no great importance, rank-order measurements will serve their purposes

Ordinal ranks, which are merely measures of relative position, are, nevertheless, frequently given specific numerical but artificial weights. Thus, degrees of attitude intensity are frequently weighted from one to five, in order that the data may be made amenable to further analysis. Such admitted approximations are very useful and are prevalently employed in social research.

Each of the four measuring devices just described possesses its individual characteristics; each follows its own rules and principles of interpretation. Furthermore, a given measure may actually exemplify two or more of the described types. For example, a direct enumeration of crimes or divorces, serving the purposes of a court docket, may at the same time serve operationally as an indirect measure of social disorganization or family instability, respectively. They all have in common, however, the important characteristic that they endeavor to express measurements in terms of symbols which permit the development and communication of social knowledge, and lay down the bases of measurement without which statistics would have no existence.

Questions and Problems

1. Define the following concepts:
   Variable
   Value
   Quantitative Variable
   Variate
   Qualitative Variable

become as elastic as would a rubber yardstick among textile merchants, each of whom would expediently adjust the unit of measure with his own amount of stretch. The acknowledged unreliability of national crime statistics is a notorious illustration of the consequences of the employment of a loosely defined concept which no amount of meticulous statistical refinement can ever overcome.

Nevertheless, concepts must still be defined as precisely as possible for purposes of sociological research; otherwise the results will be all the more ambiguous and dubious in value. Thus, in a survey on income, it is necessary to specify what constitutes income — wages, gifts, dividends, legacies, prizes — in order to obtain results which are uniform and comparable. Similarly, in a study of sentenced embezzlers, it is necessary to define the nature of this offense in order to exclude thieves who are embezzlers in name only. Social data should always be collected under the clearest possible *working definitions*, notwithstanding that such definitions will seldom be easy to formulate, and will usually contain some arbitrary elements.

(2) *Imperfections in the Measuring Instrument.* After the variable has been defined with all possible care and precision, the observations may still be inaccurate because of flaws in the measuring instrument. A faulty speedometer may overstate the mileage; a defective timepiece may run late.

Analogously, social measurements may be thrown into error by poorly conceived and inexpertly constructed measuring devices, such as the questionnaire and interview, which are so heavily relied on in sociological inquiry. An unduly long questionnaire may produce fatigue and indifference on the part of the subject, and thereby mar the accuracy of his responses. Questions may be presented in such order that successive replies are progressively distorted. Stereotyped words and phrases may evoke biassed reactions which could be avoided by the use of more neutral language.

In the face-to-face interview, imprecise and leading questions may lead to confusion and frustration and fail to elicit that sense of participation which is so essential to accurate information. In addition, the personal contact may often produce inhibitions that would not disturb the response to an anonymous ballot.

Recognizing such hazards, behavioral scientists have sought to improve their rating devices and score cards, census schedules, multiple-choice and open-end questions, focused and depth interviews, attitude and personality scales, and similar instruments in their quest for dependable data. All of these efforts testify to the principle that sound measurement is a prerequisite to statistical manipulation, which otherwise would be hollow and pretentious.

(3) *The Human Observer.* The human mind is by no means a flawless recording instrument of natural events. Even under presumably identical conditions, two equally competent individuals will obtain divergent results. Thus, it is well known that in clocking the time of a foot race, in measuring

has conferred upon mathematics the venerable prestige of being the most pure and exact of all disciplines. However, when abstract numbers are applied to concrete things, as ultimately they must, they necessarily sacrifice their purity and detachment. In the symbolism of arithmetic, for example, $4 + 4 = 8$ for every one familiar witb the conventions of the number system. However, when these numerals are made to symbolize eggs in a commercial transaction, four eggs added to four other eggs do not, in the same manner, equal eight eggs in the opinion of the economy-minded customer. The merchant struggles to approximate the exactness of pure numbers by grading the eggs according to various criteria — age, size, color — but he never succeeds completely in that endeavor. In the material world, the eight eggs are never as uniform as are the abstract units. Mere numbers will never duplicate the realities of the concrete world in which we must live.

If this fact is not borne in mind, we will make the mistake of ascribing the ideal properties of disembodied numbers to the raw objects which they represent Numbers are only a system of symbols which never completely reflect the objects they symbolize. By clothing our observations in numbers, we do not necessarily confer on the data of social science the simplicity and exactness of numbers. Observations can never be expressed without error.

However, recognition of this principle does not mean that errors should be passively accepted. On the contrary, every effort should be made to identify and reduce them, and this is possible only if the sources of these errors are known.

Sources of Measurement Error. We bere briefly consider four broad, overlapping and interlocking sources of error that bave various statistical implications: (1) vagueness in definition of the variable, (2) imperfections in the measuring instrument, (3) limitations of the human observer, and (4) the inconstancy and elusiveness of social behavior. All of these sources are likely to influence jointly any given measure; their effects cannot be readily disentangled.

(1) *Vagueness in Definition.* Measurements will be inaccurate when the variable to be measured is only vaguely defined. Divorce, nationality, war, race, family, occupation, minority, assimilation, crime, middle class, unemployment, and most other sociological concepts are difficult to define in an unambiguous manner — and all the more so over an extended duration of time, or for multiple cultural areas In general, such concepts are not physical entities which may be perceived objectively by the five senses as are apples or chemicals; they are cultural definitions with somewhat varied content which reside as social values among the individual members of a society. Scrupulous uniformity of interpretation on the part of different observers is therefore simply impossible. Such concepts sometimes

But it is as difficult to separate chance and constant errors in practice as it is easy to distinguish them in principle. Thus, it would be impossible to determine, merely by inspection, whether the socio-economic ratings made by an upper-class interviewer had a constant error, although any sociologist would be alert to that possibility. Hence, in any responsible social inquiry, it is necessary to be alert to the possible presence of constant errors, although it is their very nature to elude us.

(4) *Inconstancy and Elusiveness of Social Behavior.* When prisoners are herded by their guards into an auditorium to fill out a questionnaire, they will react otherwise than they would if each one were permitted to act spontaneously in a more conventional habitat. Persons undergoing repeated interviews by social workers, for example, are likely to become "interview-wise" and thus distort their replies by prevarication or downright untruths, either to deceive or put their best foot forward. Democrats have been known to misrepresent themselves as Republicans, according to the reports of opinion polls, when refereoce to such membership seemed more "respectable." Incomes are exaggerated because of false pride or understated out of self-interest. It is commonly known, and indeed expected, that the residents of the Elmtowns, Middletowns, and Levittowns often modify their behavior when they are conscious of being under observation. Group interaction in a laboratory setting is recognized for its artificiality; hence, to assure the normality of responses, investigators employ one-way mirrors and other comparable devices. Such sensitivity of human subjects under observation differs markedly from the relative constancy of inanimate chemicals in the laboratories of the natural sciences.

Furthermore, many forms of human behavior are indeed beyond reach in a physical sense. It is impossible to enumerate all crimes, since many remain concealed; it is impossible to observe all criminals, since many criminals remain unapprehended and out of custody. A similar argument applies to the establishment of the incidence of mental illness. By the same token, it is impossible to acquire accurate data on the "passing" between races since, at least in the racially conscious United States, its very success depends on secrecy and concealment. Thus, the acquisition of complete data on many social phenomena is effectively precluded by reason of the play of social tabus and private interests.

But social science should not therefore surrender to the comforts of an inferiority complex. Physical science, too, sometimes suffers from similar handicaps. Failure to predict the weather accurately is in part attributed to the fact that observations in remote parts of the earth and in the stratosphere are not available. Although earthquakes have been reduced to causal mechanisms, observations cannot be made in the depths of the earth which would allow geologists to predict them.

social interaction under controlled laboratory conditions, or in rating dwelling units on a socio-economic scale, there will be discrepancies ascribable to individual variability in alertness, temperament, intelligence, selective perception, and visual acuity.

When such errors compensate one another in the long run, we call them *chance errors*. There is, however, another type of error which is more difficult to control and which is the result of a tendency to record an event with a *constant error*, or *bias*. Social measures may be particularly susceptible to that type of error, since social phenomena are rarely perceived with the same degree of detachment as are physical events. An interviewer may obtain replies that conform to his own convictions, whether through misinterpretation of the replies or through suggestion to the respondent in the process of questioning. Thus, interviewers who themselves believed that the United States should keep out of World War II obtained a larger percentage of "Keep Out" opinions than did those interviewers who favored a "Help England" policy (Table 3.3 3, p. 61). Even the judges of the United States Supreme Court are swayed by personal bias and ideology, as revealed by their split decisions, and the reception that a "liberal" or "conservative" nominee to the court meets in the Senate Committee.

At times heroic efforts have been made by research agencies, necessarily with varying degrees of success, to enlist unbiased field observers and impartial investigators Thus, the Board of the Carnegie Corporation methodically searched for a disinterested scholar to direct the comprehensive study of the Negro in America which they had inaugurated in the late nineteen-thirties. The Chairman of the Board acknowledged that

> there was no lack of competent scholars in the United States who were deeply interested in the problem and had already devoted themselves to its study, but the whole question had been for nearly a hundred years so charged with emotion that it appeared wise to seek as the responsible head of the undertaking someone who could approach his task with a fresh mind, uninfluenced by traditional attitudes or by earlier conclusions, and it was therefore decided to "import" a general director. . . . And since the emotional factor affects the Negroes no less than the whites, the search was limited to countries of high intellectual and scholarly standards but with no background or traditions of imperialism which might lessen the confidence of the Negroes in the United States as to the complete impartiality of the study and the validity of its findings. Under these limitations, the obvious places to look were Switzerland and the Scandinavian countries, and the search ended in the selection of Dr. Gunnar Myrdal . . . a professor in the University of Stockholm, economic advisor to the Swedish government, and a member of the Swedish Senate.*

* From F P. Keppel's foreword to Gunnar Myrdal, *An American Dilemma*, Harper & Brothers, New York, 1944, I, pp. vi–viii.

Family
Nationality
Employment
Public opinion
Household
Delinquent

(b) Why are concepts such as these difficult to define? Compare definitions of these concepts with those in the physical sciences, e.g., molecule, acid, osmosis, alloy, earthquake, tornado.

4. Itemize some of the difficulties in making a reliable enumeration of:
The mentally ill
Illegitimate births
Interracial passers

5. Illustrate how biases might operate in measuring social distance, social class, religious tolerance, prejudice, and culture lag.

6. What are the possible errors and mistakes and their sources in the following hypothetical measures?
(a) Divorces have increased 20 per cent during the last 50 years.
(b) Twenty per cent of the labor force is unemployed.
(c) Middle-class delinquency doubled in the last decade.

7. Mistakes and errors represent deviations from the "true" value. In which instance is the true value more easily established?

Section Three

Rounded Measures

Purpose of Rounding. Because of the human equation, the imperfections in the measuring tools, and the other factors previously described, most actual measurements are destined to remain to a greater or lesser degree imprecise. The official figures on births and deaths, marriages and divorces, population, and crime can never be more than approximate. In practice, therefore, we must terminate at some convenient stage in the process of measuring the attempt at precision. In colloquial speech, we "round off" when further exactitude is neither possible nor necessary. Furthermore, rounded figures are more easily manipulated, and therefore represent a *justifiable economy of effort, even at the sacrifice of accuracy, provided* that such accuracy is not essential. Thus, we arrive at two kinds of values: *true* values, which are generally unobtainable, and *rounded* values, which are employed in actual calculation. But rounding should not be capricious; it must be systematically carried out according to established rules.

Rounding Procedures. If measures are only approximate, they must obviously be either too high or too low. When values have been rounded

Errors and Mistakes. Errors which inhere in tbe measuring process must not be confused with *mistakes.* In lay language, errors and mistakes are synonymous, but in statistical parlance these terms are clearly differentiated, even though their concrete denotations may at times be ambiguous.

A better name for mistakes would be "blunders" — blunders which derive from inexperience, incompetence, and other circumstances which a well-regulated inquiry seeks to reduce or eliminate. Although an occasional mistake is tolerated, too many of them constitute an unflattering reflection on the temperament, training, or meticulousness of the worker. A gauche interviewer who offends his subjects, the apprentice sociologist who confuses race and culture, the assistant who employs a wrong formula, produce avoidable mistakes which show up as blemishes on the countenance of the study.

However, the term "error" carries no such disparagement. Although errors, too, vary in size and seriousness, and are therefore reducible, they may be calculatingly accepted when the cost of reducing them exceeds the probable benefits of increased precision

The fundamental importance of the distinction between errors and mistakes is that there are statistical techniques which can côpe with error; but mistakes are beyond repair. In fact, chance errors are subject to the "law of error" because they are the result of chance factors and may therefore be evaluated. Even constant errors, or biases, which are the consequence of determining factors, although not analyzable according to the law of error, may nevertheless often be adjusted by appropriate statistical techniques. Understatement of age, the tendency to round at five-year intervals, the failure to report infants, and similar biased census responses may be adjusted by smoothing and other technical artifices.

But mistakes and blunders are likely to be erratic and not subject to systematic analyses. Statistical methods are therefore helpless when confronted with them; they should be banished Errors cannot be banished; consequently, their analysis and interpretation coostitute one of the central problems in statistical method.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Measurement Error
   Chance Error
   Constant Error
   Mistake

2. Distinguish between number as mere symbol and as measurement.

3. (a) In planning a sample survey, how might a statistical worker define the following concepts?

round to the nearest even number. Thus, 7.5 would be raised to 8, while 6.5 would be lowered to 6. The justification for this practice is that, theoretically, even and odd numbers occur with equal frequency in the long run. Consequently, in rounding consistently to the nearest even number, half of the values will be raised nnd the other half lowered, thereby canceling out the errors and leaving the sums free of rounding error (Columns 1 and 2). With uniform raising or lowering, however, the rounding errors would cumulate rather than compensate one another (Columns 3 and 4).

(1) *Observed Value*	(2) *Nearest Even*	(3) *Next Higher*	(4) *Next Lower*
6.5	6	7	6
4.5	4	5	4
3.5	4	4	3
7.5	8	8	7
22.0	22	24	20

Severity of Rounding. According to the United States Census, the 1950 population of Washington, D.C., was 802,178. This figure not only requires an effort to remember and manipulate, but additionally, the last several digits are not even worth such an effort since they are almost certainly unreliable, owing to difficulties in enumerating large human populations. Hence, one would not hesitate to round the figure to 802,180 or 802,200, or even to 802,000, which seems sufficiently precise for all practical purposes. With the amount of error inherent in census-taking, we could just as well have come out with a count of 801,957, which would also have rounded to 802,000. In any event, the question of how far to round is almost altogether a substantive issue whose resolution will normally be guided by two criteria: (1) our estimate of the reliability of the last digits, which, if inaccurate, should be suppressed by rounding; and (2) the degree of imprecision we are willing to tolerate. Underlying both of these criteria is, of course, the usual desire for economy of effort which rounded values permit.

Significant Digits. The digits that have been retained, on the assumption that they are reliable, are called *significant digits*. The digits of doubtful dependability are dropped. In the case of whole numbers, as in the above instance, the discarded digits are replaced by zeros merely in order to preserve the location of the decimal point. Thus, in rounding 802,178 to the nearest thousand, the vacated places must be held by zeros; otherwise the unit of count ('000) would be lost, and the meaning of the number would be lost. However, other than to indicate the unit of count, such zeros have no function. It is the significant digits that constitute the number. Thus, it is the significant digits "802" that specify how many thousands of persons reside in the city of Washington.

to the *next lower*, or last, unit, they are always too low; when rounded to the *next higher* unit, they will always be too high; and when rounded to the *nearest unit*, they will be either too high or too low. We may illustrate these differences in rounding procedures with an example of age. A person may quote his age as 20. If he follows the popular practice of giving his age as of last birthday — a practice which the United States Census follows as well — he will convey the information that his true age is somewhere between exactly 20 and not quite 21 He has accordingly rounded his age to the next lower whole number on the assumption that greater accuracy is not called for. This *rounding error*, which may be as much as a whole year, is of course not a mistake, but a known consequence of the system of rounding employed.

For actuarial purposes slightly greater precision is required. To an insurance statistician, a quoted age of 20 signifies an age between 19.5 and 20.5 years. This is rounding to the nearest whole year, or nearest birthday. It is a more exact procedure, since the rounding error in this case cannot be more than a half-year from the true age. For that reason, rounding to the nearest unit is the most common statistical procedure, and may be assumed unless it is otherwise noted.

In oriental countries, age is rounded to the next *higher* unit, but this convention is not current in the West. Nevertheless, it is adopted in certain other familiar instances For example, postage charges are assessed for "an ounce or fraction thereof," the fraction being rounded up to the full ounce. The reasons are probably: (1) quick calculation according to the first notch that tips the scales, and (2) maximization of charges. Parking garages charge for the full hour; hotels for the next full day if the guest does not vacate by check-out time. Workers who are paid by the hour, or by the day, are usually paid in full units for any fraction of the unit of time.

But whatever be the rounding procedure, the rounding unit must always be specified; one, ten, one hundred, a tenth, or a hundredth, or any other unit that satisfies our particular purpose. Infants' ages are often rounded to the week; the ages of adults may be rounded (estimated) in multiples of five years; small amounts of money are rounded to the dollar; astronomical national budgets may be rounded to the billion according to our tolerance limits.

All rounding procedures are quite uncomplicated, except when the observed value is at dead center between two adjacent values.* In this instance, an amendment to the rule is required. When rounding to the nearest unit, 7.5 could logically be rounded either to seven or eight, since both are equally proximate; but the convention in such an instance is to

* One solution is never to get stuck on dead center by the simple device of increasing the decimal accuracy of the observed measure. However, this resort will not be available when the data are obtained from secondary sources.

is the rounded number given, we would add and subtract .05 to obtain the true limits, 1.25–1.35. Thus, the true limits will always consist of one more significant digit than the rounded value.

The details of the foregoing operations hold only when rounding is to the nearest unit. When other rounding procedures are employed, the same general principles will obtain, although the computational details will vary.

Questions and Problems

1. Define the following concepts:
   True Value
   Rounded Value
   Significant Digit
   True Limits
   Midpoint
   Rounding Error
2. Round the following to the nearest whole number:
   4.03
   4.51
   4.50
   5.50
   4.501
3. Round the following (a) to the next higher, and (b) to the next lower whole number:
   5.399
   6.3
   2.25
   0.0001
4. Assume that 10 and 20 have been rounded to the nearest whole number.
   (a) Calculate:
   minimum true product
   maximum true product
   minimum true quotient of 10 ÷ 20
   maximum true quotient of 10 ÷ 20
   minimum true quotient of 20 ÷ 10
   maximum true quotient of 20 ÷ 10
   minimum true sum
   maximum true sum
   (b) What is the statistical significance of the variety in these answers?

Occasionally, even reliable digits are dropped for pure convenience, as in the informal quotation of a bank balance of $500, when the exact amount is $511.74 But the zeros used as substitutes would, of course, not be considered significant. If, on the other hand, there is reason to believe that the last zeros are exact and therefore reliable, as in the salary of $200 per week, such zeros would be regarded as significant. In such a case, the two zeros are part of the dollar count; they signify that 200 is part of the sequence 199, 200, 201, rather than 100, 200, 300.

Confidence and Precision. The population of Washington, D.C., quoted as 802,178, contains six significant digits only on the assumption that all digits are reliable and deserve confidence. A student of population, as already intimated, will question that reliability. The foregoing figure appears to be precise, but it is actually inaccurate. He will therefore fall back on the less precise but more dependable number of 802,000, since it makes no pretense of being accurate in the last three places. Paradoxically, he will have *more* confidence in the *less* precise number. Thus, we may formulate the rule: other things being equal, as precision decreases, confidence increases; as precision increases, confidence decreases. A little reflection and experience will support the simple logic of this generalization, which is employed in many statistical situations.

True Limits of Rounded Numbers It is obvious that in rounding to the nearest whole unit, observed values such as 4.7, 4.9, 5.2, and 5.4 will all be rounded to 5. Hence, when we encounter that rounded value, we must be aware that the observed value, inaccurately represented by 5, may lie anywhere within the interval, 4.5 to 5.5. Such a reconstruction of what are labeled the *true limits* from the rounded number is the reverse of the process of rounding That is, we infer that the original observed value must have been at a point somewhere along the interval, extending a half unit on either side of the rounded value, 5 Therefore, the rounded value is the *midpoint* of that interval and lies exactly half-way between the true limits.

However, these true limits will be impossible to establish when the severity of the prior rounding is not indicated, as may be seen in the following illustration:

Rounded Number	*Unit of Measure*	*True Limits*
400	Ones	399.5–400.5
400	Tens	395 –405
400	Hundreds	350 –450

From these examples, we may formulate a working rule for establishing the true limits as follows: *add to, and subtract from, the rounded number one-half of the given unit of measure.* If ten was the rounding unit, we would add and subtract 5, resulting in the true limits of 395–405. If 1.3

Grouping of Data 3

Section One

Quantitative Variables

All statistical knowledge is knowledge of large aggregates. But all collection of data must begin with the individual case. This is an important principle that is sometimes lost sight of after the individual item has been swallowed up by the mass. Nevertheless, we do not linger on the individual event, for reliable knowledge rests on the repetition of experience. But the human mentality, unaided, cannot encompass a multitude of sepamte observations. If it attempts to do so, it is likely to be unduly impressed by conspicuous cases or influenced by expectations and desires, so that conclusions will be distorted and misleading It is one of the functions of statistics to reduce such hazards by its techniques of organizing and condensing masses of data into comprehensible form. So fundamental is this function that statistics has frequently been characterized as the "science of large numbers."

There are two basic types of statistical groupings, corresponding to the previous classification of variables: (1) the quantitative, in which variates are grouped into ordered class intervals; and (2) the qualitative, in which attributes are distributed into disparate categories. Thus, we may group the incomes of 1,000 persons into any convenient number of graduated intervals, but the classification of the same persons by sex is necessarily twofold, and the order of categories is arbitrary. In spite of their differences, both quantitative and qualitative classifications culminate in the *frequency distribution*, which is simply a tally of the cases in each of the respective classes.

A possible exception to the foregoing division is the *time series*, which is an ordering of values according ta an entirely different dimension, namely, chronology of occurrence.

6. If a girl's height to the nearest inch is 65, between what true limits would her observed height fall?

7. An array of homicide rates extends from .5 to 17.6, rounded to the nearest tenth. What are the true limits of this interval?

8. Designate the significant digits in each of the following numbers:

   2.3
   0 0203
   .600
   .006
   800
   8,000

9. Distinguish between .65 and .650

10. (a) Is 51 centimeters necessarily less precise than .51 meter?
    (b) Is 5,280 feet more precise than one mile?
    (c) Is 50 miles less precise than 50.25 miles?

11 Two baseball players have 60 and 59 hits out of 190 times at bat, respectively. How would one determine the number of decimal places in the batting average? Why does it usually consist of three places? [*Note*· The "batting averaga" is defined as the number of hits divided by the number of times at bat.]

12 Explain why rounding to the nearest unit introduces less rounding error than rounding to the last unit.

13. (a) Formulate the rule for calculating the true limits of a number rounded to the last unit. (b) What is the maximum error in such rounding; and in what direction does it lie from the rounded value?

Selected References

Bureau of the Census, *Census of Population: 1950.* Volume 2, Part 1, *United States Summary* U.S. Government Printing Office, Washington, D.C., 1953. Pages 1–66.

Cohen, Morris, and Ernest Nagel, *An Introduction to Logic and Scientific Method.* Harcourt, Brace and Company, New York, 1934. Chapter 15

Dantzig, Tobias, *Number The Language of Science.* Fourth edition. Doubleday Anchor Books, Garden City, N.Y., 1956.

Lazarsfeld, Paul F., "Problems in Methodology," in *Sociology Today.* Edited by Robert K. Merton, Leonard S. Cottrell, Jr. Basic Books, Inc., New York, 1959. Pages 39–78

Torgerson, Warren S , *Theory and Methods of Scaling* John Wiley & Sons, Inc , New York, 1958 Chapters 1 and 2.

Walker, Helen M., *Mathematics Essential for Elementary Statistics.* Revised edition. Henry Holt and Company, New York, 1951. Chapters 5 and 6.

Table 3.1.1a (*Continued*)

City	(1) Population	(2) Suicides	(3) Rate per 100,000
Houston, Texas	596,163	88	14.8
Indianapolis, Indiana	427,173	66	15.4
Jacksonville, Florida	204,517	22	10.8
Jersey City, New Jersey	299,017	20	6.7
Kansas City, Kansas	129,553	14	10.8
Kansas City, Missouri	456,622	50	10.9
Knoxville, Tenn.	124,769	13	10.4
Little Rock, Ark.	102,213	15	14.7
Long Beach, California	250,767	61	24.3
Los Angeles, California	1,970,358	371	18.8
Louisville, Kentucky	369,129	37	10.0
Memphis, Tenn.	396,000	26	6.6
Miami, Florida	249,276	34	13.6
Milwaukee, Wisconsin	637,392	74	11.6
Minneapolis, Minn.	521,718	76	14.6
Mobile, Ala.	129,009	9	7.0
Montgomery, Ala.	106,525	8	7.5
Nashville, Tenn.	174,307	23	13.2
New Bedford, Mass.	109,189	13	11.9
New Haven, Conn.	164,443	19	11.6
New Orleans, La.	579,445	40	7.0
New York City, New York	7,892,957	837	10.6
Newark, New Jersey	438,776	62	14.1
Niagara Falls, N.Y.	191,555	5	2.6
Norfolk, Virginia	213,513	24	11.2
Oakland, California	384,575	58	15.1
Oklahoma City, Okla.	243,504	22	9.0
Omaha, Nebraska	251,117	31	12.3
Pasadena, Calif.	104,577	21	20.1
Paterson, New Jersey	139,336	19	13.6
Peoria, Illinois	111,856	15	13.4
Philadelphia, Pa.	2,071,605	178	8.6
Phoenix, Arizona	106,818	25	23.4
Pittsburgh, Pa.	676,806	78	11.5
Portland, Oregon	373,628	66	17.7
Providence, Rhode Island	248,674	24	9.6
Reading, Pa.	109,320	18	16.5
Richmond, Virginia	230,310	42	18.2
Rochester, New York	332,488	47	14.1

(*Table continued*)

Table 3.1.1a *Suicides and Suicide Rates, 107 Large U.S. Cities, 1950*

City	(1) Population	(2) Suicides	(3) Rate per 100,000
Akron, Ohio	274,605	34	12.4
Albany, N.Y	134,995	11	8.1
Allentown, Pa.	106,756	8	7.5
Atlanta, Ga.	331,314	37	11.2
Austin, Texas	132,459	8	6.0
Baltimore, Maryland	946,708	101	10.6
Baton Rouge, La.	125,629	7	5.6
Berkeley, California	113,805	24	21.1
Birmingham, Alabama	326,037	15	4.6
Boston, Mass.	801,444	84	10.5
Bridgeport, Conn.	158,709	14	8.9
Buffalo, New York.	580,132	53	9.1
Cambridge, Mass.	120,740	7	5.8
Camden, New Jersey	124,555	13	10.4
Canton, Ohio	116,912	7	6.0
Charlotte, N.C.	134,042	11	8.2
Chattanooga, Tenn.	131,041	6	4.6
Chicago, Illinois	3,620,062	407	11.2
Cincinnati, Ohio	503,998	71	14.1
Cleveland, Ohio	914,808	106	11.6
Columbus, Ohio	375,901	36	9.6
Corpus Christi, Texas	108,485	15	13.8
Dallas, Texas	434,462	33	7.6
Dayton, Ohio	243,872	28	11.5
Denver, Colorado	415,786	81	19.5
Des Moines, Iowa	177,965	30	16.8
Detroit, Michigan	1,849,568	185	10 0
Duluth, Minnesota	104,511	16	15.3
Elizabeth, New Jersey	112,817	21	18.6
El Paso, Texas	130,485	11	8.4
Erie, Pa.	130,803	19	14.5
Evansville, Indiana.	128,636	21	16 3
Fall River, Mass.	111,963	8	7.1
Flint, Michigen	163,143	15	9.1
Fort Wayne, Indiana	133,607	11	8.2
Fort Worth, Texas	278,778	28	10.0
Gary, Indiana	133,911	20	14.9
Grand Rapids, Mich.	176,515	19	10.8
Hartford, Conn.	177,397	17	9.6

(*Table continued*)

procedure in the grouping of quantitative data, we shall analyze the suicide rates of 107 Americao cities of 100,000 populatioo aod over for 1950.

Table 3.1.1a preseots first the actual number of suicides in each of the 107 major cities io the United States. It is evident that such absolute numbers are oot very revealing. Not unexpectedly, the larger cities have a larger oumber of suicides, and the smaller cities have a smaller volume. Clearly, the data are in a rather elementary stage of compilatioo.*

Io order to infuse meaning into the raw numbers, it is oecessary to relate the number of suicides to the size of the population of the respective cities. Only theo will the differential teodency of the cities toward suicide become appareot. This, of course, requires the calculatioo of a rate, which has the effect of norming, or staodardizing, all eities to the same population size, in this instaoce, 100,000. We thereby eliminate the variahle of population size, which has now become a constant. Because of the small number of suicides in relation to the general population, the rate is computed on the basis of 100,000, thereby reducing the excessive decimal places, which are difficult to read and less quiekly grasped.

The individual suicide rates of these eities (Column 3) have already been reduced to some order hy an alphabetical arrangement, which is useful for many purposes. But the alphabetical position of a city rests on a quite arhitrary system which is totally unrelated to the magnitude of the rates. It is therefore not a statistical order; it is merely a cataloguing device for the ready location of the individual city in which one may be momentarily interested. As far as a general impression of the pattern of distribution of suicide rates is concerned, the cities might just as well he listed at random.

A first step in the direction of putting the materials in orderly form is to arrange the rates in order of magnitude, from the lowest to highest, or the reverse. This is called ao *array*, and is comparahle to lining up a company of soldiers according to height, or sortiog apples according to size. We have now lost both the population size aod separate ideotity of the cities, and we are beginniog to conceive of these values as ordered variates — which is essentially a statistical conception.

This new view of the data (Table 3.1.1b), afforded hy the arrangement io order of magnitude, renders a clearer picture of the concentration of values in the middle of the array. Indeed, the differences between any of these values and its neighbors are usually so slight that it would seem pedantic to assign great significance to what seems to he mere chaoce variation. No sociological distinctions between cities could be drawn from

* Although from the standpoint of grouping the data in this table are in a rather elementary stage, actually we should realize that, from the standpoint of the process of collection of data, they are already in an advanced stage of compilation. The original sources are the death certificates, which are collected by the offices of vital statistics in the individual cities. These have been assembled, summarized, and then reported to the state and national offices of vital statistics, and finally published as descriptive tables by the Federal Government. It is from these published tables that this material has been extracted.

Table 3.1.1a (*Concluded*)

City	(1) Population	(2) Suicides	(3) Rate per 100,000
Sacramento, California	137,572	36	26.1
St. Louis, Missouri	856,796	90	10.5
St. Paul, Minn.	311,349	30	9.6
Salt Lake City, Utah	182,121	23	12.6
San Antonio, Texas	408,442	48	11.8
San Diego, Calif.	334,387	70	21.0
San Francisco, Calif.	775,357	223	28.8
Savannah, Georgia	119,638	20	16.7
Scranton, Pa.	125,536	6	4.8
Seattle, Washington	467,591	124	26.5
Shreveport, La.	127,206	11	8.6
Somerville, Mass.	102,351	6	5.9
South Bend, Indiana	115,911	26	22.4
Spokane, Wash.	161,721	27	16.7
Springfield, Mass.	162,399	8	4.9
Syracuse, New York	220,583	22	10.0
Tacoma, Washington	143,673	39	27.1
Tampa, Florida	124,681	31	24.9
Toledo, Ohio	306,616	46	16.1
Trenton, New Jersey	128,009	24	18.8
Tulsa, Oklahoma	182,740	22	12.0
Utica, New York	101,531	13	12.8
Washington, D.C.	802,178	92	11.5
Waterbury, Conn.	104,477	9	8.6
Wichita, Kansas	168,279	24	14.3
Wilmington, Del.	110,356	13	11.8
Worcester, Mass.	203,486	11	5.4
Yonkers, New York	152,798	17	11.1
Youngstown, Ohio	168,330	13	7.7

Sources U.S. Bureau of the Census. *U.S. Census of the Population: 1950* Vol. I, *Number of Inhabitants*, U.S. Government Printing Office, Washington, D.C., 1952; U.S. National Office of Vital Statistics, *Vital Statistics of the U.S. 1950*, Vol. III, U.S. Government Printing Office, Washington, D.C., 1953.

Grouping Procedure. Since age, weight, income, and various types of rates, are things of which there may be more or less, they are quantitative in nature. Cities may vary in amount of suicide, crime, or divorce; individuals may vary in income, weight, or age; marriages may vary in duration and happiness. But such variation is not random and chaotic; rather it usually conforms to some specific model which will emerge only after the individual items have been judiciously grouped. To illustrate the

the class intervals. In this respect, there are no rigid prescriptions, although obviously the ungrouped array must be sufficiently compressed to achieve the purpose of grouping. Broadly speaking, grouping should be adapted (1) to the nature of the data, (2) to the objectives which the classification intends to serve, and (3) to the background and needs of the readers for whom it is intended. Thus, data embracing an immense range, such as the ages of the United States population, would require a different grouping than would the ages of American school children. The oversimplified statistics in the popular brochures of insurance companies, depicting the relative incidence of heart disease and cancer, may be very instructive for the lay public, but inadequate for scientific uses. A grouping of data intended for a professional audience could be more intricate than one planned for popular consumption.

In spite of this flexibility, it is customary to lay down a few technical rules which could normally serve as a guide unless specific conditions dictate otherwise.

(1) The *number* of class intervals should ordinarily be no more than 15, nor less than 10. Fifteen intervals are usually sufficient to reveal the pattern of distribution, and yet not so numerous that it cannot be readily apprehended. On the other hand, if fewer than ten intervals are employed, the salient features of the distribution may be obscured.

(2) The *size* of the interval should be a whole number, and, whenever practicable, of convenient divisibility such as 2, 10, or 25. Additionally, there may be some practical advantage attached to the use of multiples of 10, which is consistent with our decimal system. With units in feet and inches, minutes and hours, the requirement of adhering to the decimal system may of course be waived.

(3) Intervals should, whenever possible, be of *uniform width*. They are then most readily grasped, and greatly facilitate further computation. However, the rule of equal intervals may at times have to give way in the face of special circumstances. Thus, the school population may call for age intervals of 6–13, 14–17, and 18–22, otherwise information essential to the school administrator would be concealed. Similarly, in the classification of cities by size of population, or families by income brackets, uniformity of class intervals cannot be maintained. At the lower end of the range, variates are numerous and differences are small, while in the upper brackets, cases are infrequent and gaps are large. Hence, class intervals should increase progressively in size throughout the distribution. However, whenever unequal intervals are used, the larger intervals should be multiples of the smaller ones, especially when graphs are contemplated.

(4) The ends of a frequency distribution may be either *open* or *closed*. Although the ends are normally closed, it is often convenient to leave one or both extremes *open*, as in the National Office of Vital Statistics tabulation of births by age of mother:

Table 3.1.1b *Array of Original Suicide Rates*

2.6	6.7	8.6	10.0	11.1	11.8	14.1	15.4	20.1
4.6	7.0	8 6	10.0	11.2	11.9	14.1	16.3	21.0
4.6	7.0	8.6	10.0	11.2	12.0	14.1	16.5	21.1
4.8	7.1	8.9	10.4	11.2	12.3	14.3	16.7	22.4
4.9	7.5	9.0	10.4	11.2	12.4	14.5	16.7	23.4
5.4	7.5	9.1	10.5	11.5	12.6	14.6	16.8	24.3
5.6	7.6	9.1	10.5	11.5	12.8	14.7	17.7	24.9
5.8	7.7	9 6	10 6	11.5	13.2	14.8	18.2	26.1
5.9	8.1	9.6	10.8	11.6	13.4	14.9	18.6	26.5
6.0	8.2	9.6	10.8	11.6	13.6	15.1	18.8	27.1
6.0	8 2	9 6	10 8	11.6	13.6	15.1	18.8	28.8
6.6	8 4	10.0	10.9	11.8	13.8	15.3	19.5	

the numerous slight variations existing within the list. Not only are such differences meaningless, but they actually clutter our view of the underlying pattern of variation. Therefore, why not slough them off? Rounding the numbers would of course reduce the trivial differences between the adjacent values (Table 3.1.1c).

Table 3.1.1c *Array of Suicide Rates, Rounded to Nearest Whole*

3	7	9	10	11	12	14	15	20
5	7	9	10	11	12	14	16	21
5	7	9	10	11	12	14	16	21
5	7	9	10	11	12	14	17	22
5	8	9	10	11	12	14	17	23
5	8	9	10	12	13	15	17	24
6	8	9	10	12	13	15	18	25
6	8	10	11	12	13	15	18	26
6	8	10	11	12	13	15	19	26
6	8	10	11	12	14	15	19	27
6	8	10	11	12	14	15	19	29
7	8	10	11	12	14	15	20	

But this step is not yet sufficient to eliminate all the confusion which usually results from a large number of items. Furthermore, if one assumes that many other arrays are still more bulky and complex than this one, it is obvious that some device must be made available to render this array comprehensible to the social analyst. This device consists in *grouping* identical or contiguous values into convenient *class intervals.* The form resulting from this procedure goes by the name of *frequency distribution.*

Number and Size of Class Intervals. Before we can proceed to the actual tabulation, we must first determine the appropriate number and size of there are no vacant intervals; and the outline of the distribution assumes the shape of a curve with a single peak in the middle. The suicide rates now lend themselves to a quick description as clustering around 11 or 12 per 100,000 population, with variations below and above this figure.

In this table, the values have been listed from low to high. This practice, which is that of the United States Census and prevails generally in the tabulation of social statistics, differs from the convention employed by psychologista and educators who reverse the order of values. The difference is, of course, superficial rather than essential. In the tabulation of test scores and school marks it is perhaps more natural to begin with the highest score, against which all other scores are judged. In general, however, with such variables as age and income, one is likely to begin counting from the zero origin, in conformity with habits acquired in childhood and reinforced through daily use.

Any frequency tabulation may at first produce the impression that we are becoming more and more inexact in an endeavor to condense the data to a more manageable level. It is true, every condensation does some violence to detail. However, in view of the fact that the foregoing suicide rates are for a specific year, and that these rates in any event will vary somewhat in the same city from year to year, there would be no purpose in insisting on such exaggerated exactness. We have merely discarded detail which is probably of no use anyway. Grouping is therefore not only more convenient, but is fully justified and actually necessary in order to display the fundamental pattern of events. If, in the interest of maneuverability, however, grouping is too broad and coarse, *grouping errors* may occur which must later be taken in account in the interpretation of the materials.

Rounded and True Limits.[*] The class limits of the frequency distribution (Table 3.1.1d) have been designated by rounded numbers: 3–5, 6–8, and so on. However, according to prevailing convention, a rounded number is itself considered the midpoint of some definite interval: 3 is the midpoint of the interval 2.5–3.5, and 5 is the midpoint of the interval 4.5–5.5. It follows that the *true limits* of the first and successive class intervals would be written as follows:

2.5– 5.5
5.5– 8.5
8.5–11.5
. . . .
. .
26.5–29.5

[*] Some texts employ the terms "written" and "true" limits where this text designates them as "rounded" and "true." Since "true" limits can be, and frequently are, written down, and since the "written" limits are actually rounded limits, it would seem efficient to call "rounded" values by name.

Age of Mother
Under 15 years
15–19 years
20–24 years
25–29 years
30–34 years
35–39 years
40–44 years
45 years and over

Such *open ends* are resorted to in order to avoid the necessity of itemizing small, irregular and essentially trivial frequencies which extend well beyond the limits of the significant range of the data. However, for technical reasons, it is a good rule to close the ends whenever possible. Otherwise subsequent calculations, such as the arithmetic average, are impossible, and graphing becomes awkward.

Tabulation. In grouping the suicide rates, the first step is to divide the range into an effective number of class intervals having a convenient width. The range of the array extends from 3 to 29, a spread of 27 points. A little exploration will reveal that an interval of five would yield too few groups to provide the desired discrimination, while an interval of only one point would obviously not produce the desired condensation. We therefore select a 3-unit interval as the optimum. Then, starting with a multiple of 3 as the lower limit, we write down the class limits and proceed with the tally as shown in Table 3.1.1d.

Table 3.1.1d *Frequency Tally of Suicide Rates*

ROUNDED CLASS LIMITS (*X*)	TALLY	FREQUENCY (*f*)
3– 5	𝍸 /	6
6– 8	𝍸 𝍸 𝍸 ///	18
9–11	𝍸 𝍸 𝍸 𝍸 𝍸 ////	29
12–14	𝍸 𝍸 𝍸 𝍸 ////	24
15–17	𝍸 𝍸 ///	13
18–20	𝍸 //	7
21–23	////	4
24–26	////	4
27–29	//	2
		$N = 107$

This frequency distribution of suicide rates commends itself to common statistical sense. The frequencies are neither too large nor too small;

Age of Mother
Under 15 years
15–19 years
20–24 years
25–29 years
30–34 years
35–39 years
40–44 years
45 years and over

Such *open ends* are resorted to in order to avoid the necessity of itemizing small, irregular and essentially trivial frequencies which extend well beyond the limits of the significant range of the data. However, for technical reasons, it is a good rule to close the ends whenever possible. Otherwise subsequent calculations, such as the arithmetic average, are impossible, and graphing becomes awkward.

Tabulation. In grouping the suicide rates, the first step is to divide the range into an effective number of class intervals having a convenient width. The range of the array extends from 3 to 29, a spread of 27 points. A little exploration will reveal that an interval of five would yield too few groups to provide the desired discrimination, while an interval of only one point would obviously not produce the desired condensation. We therefore select a 3-unit interval as the optimum. Then, starting with a multiple of 3 as the lower limit, we write down the class limits and proceed with the tally as shown in Table 3.1.1d.

Table 3.1.1d *Frequency Tally of Suicide Rates*

ROUNDED CLASS LIMITS (X)	TALLY	FREQUENCY (f)
3– 5	𝍸 /	6
6– 8	𝍸 𝍸 𝍸 ///	18
9–11	𝍸 𝍸 𝍸 𝍸 𝍸 ////	29
12–14	𝍸 𝍸 𝍸 𝍸 ////	24
15–17	𝍸 𝍸 ///	13
18–20	𝍸 //	7
21–23	////	4
24–26	////	4
27–29	//	2
		$N = 107$

This frequency distribution of suicide rates commends itself to common statistical sense. The frequencies are neither too large nor too small;

higher interval. For example, observed 6.0 becomes "rounded" 6 and hence joins the interval 6–8.

Midpoint of Class Interval. Since it is impossible to subject a class interval to further calculation, it is often necessary to designate a single value for all the items which are included within its boundaries. The selection of this value must be guided by the principle of representativeness as well as expediency. On the assumption that the items are more or less uniformly distributed between the class boundaries, a point half way between the true boundaries would satisfy our needs. Accordingly, we select the *midpoint* of the true interval to represent all the values within the interval. In effect, we round off all individual values to the interval midpoint. But by that maneuver, we introduce grouping error, owing to the fact that few if any of the observed items within the interval will fall exactly at the midpoint. When items are evenly distributed, the errors produced by arbitrary grouping will offset one another; however, when the items are bunched in one end of the interval, the errors will not balance one another and the cumulated grouping error will infect subsequent calculations such as the mean. Hence, *grouping error* is defined as the amount by which the average value of a set of grouped data diverges from the average value of the same data in ungrouped form. When there is reason to believe that grouping error is severe, correction formulas should be applied so that averages derived from grouped data will more closely approximate those of the original, ungrouped data.

To find the midpoint of any interval, we divide its true width by two, and add the result to the true lower limit. An equivalent arithmetic procedure is to add directly the values of the true limits, and divide by two. The procedure is applicable irrespective of the manner of rounding or the type of data (whether continuous or discrete). To illustrate, we present in Table 3.1.2 a set of rounded limits (3–5, 6–8), whose true limits and midpoints vary, however, according to the prior rounding procedure.

Table 3.1.2 *Determination of Interval Midpoint, Rounding to Nearest Whole Number and Last Whole Number*

ROUNDING PROCEDURE	ROUNDED LIMITS	TRUE CLASS LIMITS	CLASS WIDTH	ONE-HALF CLASS WIDTH	INTERVAL MIDPOINT
Nearest whole number	3–5	2.5–5.5	3	1.5	4
	6–8	5.5–8.5	3	1.5	7
Last whole number	3–5	3.0–6.0	3	1.5	4.5
	6–8	6.0–9.0	3	1.5	7.5

It is evident that the true limits, which are of course spaceless points, constitute the common boundaries between adjoining class intervals and thereby assure the continuity of the scale, as in Figure 3.1.1.

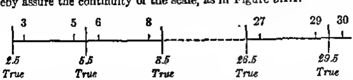


FIGURE 3 1.1 *Rounded and True Limits*

At first glance, the common boundaries might seem to confuse the task of grouping In what interval, for example, should we locate an observed value of 5.5, which falls right on the boundary of contiguous class intervals? The resolution of this dilemma has already been offered: rounding to the nearest even value, we would place it in the interval 6–8.

It will become increasingly apparent that the true limits are the only limits with which we can do any statistical business. We may accordingly raise the question: why carry along the rounded limits in the vocabulary baggage? The answer is threefold: (1) rounded limits are easier to read, whereas the more detailed but accurate true limits "clutter up" the table visually; (2) rounded limits are adequate for tallying purposes; and (3) they can always be transformed into true limits whenever necessary.

In order to reconstruct the true limits from the rounded, it is of course necessary to know the prior rounding procedure: whether to the nearest or the next lower unit, which is not evident from the outward appearance of the rounded numbers. From the appearance of the following rounded class limits, for example, it is not evident whether the suicide rates were rounded to the nearest or last whole unit:

3– 5
6– 8
9–11

Had they been rounded to the nearest whole number, the *true limits* would be written·

2.5– 5.5
5.5– 8.5
8.5–11.5

But had they been rounded to the last whole unit, the true limits would read:

3.0– 6 0
6.0– 9.0
9.0–12.0

Where rounding is to the last unit, and where the observed value falls on a boundary, we "drop the zero" and thereby place the item in the next

Technique of Table Construction. We must always anticipate that statistical materials will ultimately be presented in orderly tabular form to some audience. In order to ensure maximum intelligibility and effectiveness in display, it is necessary that such presentation conform to well-established conventions that prescribe the structure and content of the table. As it stands, Table 3.1.1d is still a working table, and not yet wholly presentable to its public. It must still be refined in several respects: a substantive title must be attached, the temporary tally marks removed, and clear captions provided.

The *title* must convey in plain manner the content of the table; therefore, the first term of the title should normally designate the subject matter of the table. Then should follow as needed, in order of priority, an indication of the classification of the materials, the geographical area and the period of time to which they pertain. To express it patly: the elements of the title should be arranged in order of the "what, how, where, and when" of the tabular information.

Architecturally, the table consists of three parts: *stub, caption,* and *body.* The stub, located in the left sector, where the Occidental eye usually looks first, exhibits the primary classification of the materials. It should always be identified by its own heading. The body, to the right of the stub, comprises the frequencies, inserted in the cells created by the intersections of the rows and columns. The row frequencies are identified by the stub, while the column frequencies are designated by the caption, which consists of the one or more secondary classifications with appropriate labels. To illustrate this conventional arrangement of parts, we reproduce in Table 3.1.4 a typical United States Census table.

Table 3.1.4 *School Enrollment Rates by Age, Selected Years, 1910–1957*

	Year	Age					
		5–6	7–13	14–17	18–19	20	Caption
Stub	1957	78.6	99.5	89.5	34.9	..	
	1950	39.3	95.7	83.9	32.3	17.9	
	1940	43.0	95.0	79.3	28.9	12.5	Body
	1930	43.2	95.3	73.1	25.4	13.1	
	1920	41.0	90.6	61.6	17.8	8.3	
	1910	34.6	86.1	58.9	18.7	8.4	

Source: Donald J. Bogue, *The Population of the United States*, The Free Press, Glencoe, Ill., 1959, p. 329 (adapted).

Relative Frequencies. It is only by way of comparison with other distributions that a given frequency distribution takes on meaning. But it is

Since the rounded values are indistinguishable for the two types of rounding, care must be exercised in reading correctly the rounded value. If, for example, rounded 3 is mistakenly read as 3.0–4.0, instead of 2.5–3.5, the midpoint would be located one-half unit too high.

Discrete values follow the same rules of grouping as are applied to continuous data, except for a minor technical adaptation. Since a discrete variable jumps from one integer to the next instead of moving continuously over a specified range, a discrete variate is and remains a whole number. In certain situations, however, we must make the factitious assumption that a discrete variable is after all continuous and that the observed values have been rounded to the nearest whole. True limits and class midpoints calculated on this assumption are illustrated in Table 3.1.3.

Table 3.1.3 *Number of Dwelling Units per Block, Los Angeles, 1940*

Rounded Limits	True Limits	Class Midpoint	Frequency (Per Cent)
0– 4	..... .*	2	25.44%
5– 9	4.5– 9.5	7	9.52
10– 14	9.5–14.5	12	8.60
15– 19	.. . ..	17	7.76
20– ·24	......	....	7.62
25– 29	... ...	....	7.01
30– 39	... ...	...	10 66
40– 49	....	.. .	6.42
50– 59	.. . ..	....	4.47
60– 69	. ....	64.5	2.95
70– 79	....	74.5	2.04
80– 89	. . ...	....	1.73
90– 99	.. ....	...	1.29
100–149	.. .. .	....	2.95
150–199	....	....	.86
200–249	. . ..	....	.36
250–299	.. ..	. ..	.16
300–399	... ..	. ..	.11
400+	.. .	. ..	.05
Total			100 00%

* Missing values to be calculated by student.

Source: U.S. Bureau of the Census, *U S. Census of Population and Housing· 1940, Supplement to the First Series, Housing Bulletin for California, Los Angeles Block Statistics*, U.S. Government Printing Office, Washington, D C.

Table 3.1.6 *Cumulative Age Distribution, U.S. Population, 1950*

Age *	Per Cent	"Less Than" Cumulative Per Cent	"Or More" Cumulative Per Cent
0– 4	10.7	10.7 (less than 5)	100.0 (0 or more)
5– 9	8.8	19.5 (less than 10)	89.3 (5 or more)
10–14	7.4	26.9	80.5
15–19	7.0	33.9	73.1
20–24	7.6	41.5	66.1
25–29	8.1	49.6	58.5
30–34	7.6	57.2	50.4
35–39	7.5	64.7	42.8
40–44	6.8	71.5	35.3
45–49	6.0	77.5	28.5
50–54	5.5	83.0	22.5
55–59	4.8	87.8	17.0
60–64	4.0	91.8	12.2
65–69	3.3	95.1	8.2
70–74	2.3	97.4	4.9
75 and over	2.6	100.0 (less than 100)	2.6 (75 or more)

* Age rounded to the last whole year.

Source: U.S. Bureau of the Census, *U.S. Census of the Population: 1950*, Vol. II, *Characteristics of the Population*, Part I, *United States Summary*, U.S. Government Printing Office, Washington, D.C., 1953.

population is less than 5 years of age; to find what percentage of the population is less than 10 years of age, we cumulate the first *two* class frequencies: 10.7% + 8.8% = 19.5%. All successive cumulative frequencies are similarly obtained. The highest interval, however, offers a slight difficulty, since it has no specified upper limit — it is "open ended." In fact, the United States Census finds the scatter of items in the higher ages so thin that it has already cumulated the last several intervals on an "or more" basis. To provide an upper limit for the "less than" cumulation, it is permissible in this instance to close the interval arbitrarily at 100 years.

The "or more" cumulation is read from the true lower limit of the intervals. Thus, 2.6 per cent of the population is 75 or more. In order to obtain the per cent of population 70 or more, we cumulate the frequencies of the two highest intervals: 2.6% + 2.3% = 4.9%; and so on to the 100 per cent cumulation.

From these two tables, any number of useful readings may be made: 8.2 per cent of the population is 65 years and over; 49.6 per cent of the population is under 30 years of age. Comparable questions could be answered from the appropriate tabulations: what proportion of families have incomes of $10,000 and over? what proportion of families have six or more children? what proportion of unemployed persons have been without work for 6 months or more?

difficult to compare two or more different distributions that have widely varying totals. To norm them, or make them comparable, the absolute distributions must be reduced to relative, or percentage distributions. The conversion of an absolute to a percentage frequency is accomplished by dividing the class frequency by the total, and multiplying the result by 100 — a familiar arithmetic procedure. Table 3.1.5 illustrates the convenience of this method of comparison.

Table 3.1.5 *U.S. Population, by Age, 1880 and 1950*

Age	Number		Per Cent	
	1880	1950	1880	1950
0– 9	13,394,176	29,363,256	26.7%	19.5%
10–19	10,726,601	21,735,866	21.4	14.4
20–29	9,168,393	23,724,088	18.2	15.7
30–39	6,369,362	22,763,393	12.7	15.1
40–49	4,558,256	19,274,438	9.2	12.8
50–64	4,215,536	21,566,783	8.4	14.3
65+	1,723,459	12,269,537	3.4	8.2
Total	50,155,783	150,697,361	100.0%	100.0%

Source: U.S. Bureau of the Census. *U.S. Census of the Population: 1950*, Vol. II, *Characteristics of the Population*, Part I, *United States Summary*. U.S. Government Printing Office, Washington, D.C., 1953.

Cumulative Distribution. At times we may simply want to know the number or percentage of items above or below a certain cutting point on the scale. A welfare administrator, for example, will want to know how many persons are 65 years or older. Such results may be readily obtained from the *cumulative frequency distribution*, which is derived from the *simple frequency distribution* by a process of merging successive class intervals until all have been cumulated. Such a procedure reveals the frequency of items below and above each class boundary, and thus permits the cutting point to be easily moved to any convenient position according to the interests and needs of the reader. Table 3.1.6 — a work table — demonstrates how the cumulated tabulation is derived from the simple serial table.

In cumulating, it is possible to employ as an origin either the lowest or highest class interval and proceed to the opposite end. When the lowest interval is employed as an origin, we obtain what may be conversationally termed a "less than this much" cumulative distribution; when the highest is employed, we obtain a "this much or more" cumulative frequency distribution.

The "less than" cumulation supplies the per cent of population below the true upper limit of each class interval. Thus, 10.7 per cent of the

Table 3.1.7 *Delinquency Rates, 140 Local Areas, Chicago*

0.8	1.6	1.8	2.5	18.2	5.2	0.6
0.8	1.9	2.9	3.7	4.9	5.1	0.6
1.5	1.5	4.6	5.5	2.7	1.7	2.1
2.6	0.9	3.5	7.4	2.3	1.5	2.2
1.0	2.4	3.0	11.8	5.8	2.6	3.7
0.7	1.7	1.9	12.3	8.8	4.0	1.0
0.5	2.2	4.4	12.1	18.9	5.8	3.9
1.1	2.2	2.9	3.0	2.3	9.4	0.6
0.8	2.3	3.4	4.3	2.7	2.4	1.9
0.8	4.3	4.8	5.0	3.2	2.2	4.6
1.4	1.6	6.1	5.7	7.0	3.5	1.5
1.9	1.2	7.8	11.9	13.4	2.8	1.3
2.8	2.1	2.7	9.5	17.5	2.2	1.3
1.0	1.9	3.1	14.8	4.5	2.5	6.0
1.1	2.2	5.2	2.5	2.7	3.9	4.2
0.9	2.9	9.0	2.1	3.4	1.5	4.2
0.7	2.5	11.4	5.0	4.5	0.8	2.0
1.6	4.2	9.5	5.1	9.4	0.6	2.8
3.1	3.1	12.1	3.7	14.8	1.2	2.1
2.1	1.9	2.7	5.0	1.6	3.0	2.5

Source: Clifford R. Shaw and Henry D. McKay, *Juvenile Delinquency and Urban Areas*, The University of Chicago Press, Chicago, 1942, p. 53.

12. Calculate the width and midpoint of each of the following intervals, assuming that data are rounded to (a) the nearest unit and (b) to the next lower unit:

6– 8	2– 5	0– 4
9–11	6– 9	5– 9
12–14	10–13	10–14

13. From the given class limits and midpoints, determine:
    (a) whether class limits are *true* or *rounded*
    (b) the size of the class interval
    (c) the manner of rounding, if any.

Limits	*Midpoint*	*Limits*	*Midpoint*
20–24	22.5	0– 9	4.5
25–29	27.5	10–19	14.5
14.5–19.5	17.0	.5–2.4	1.45
19.5–24.5	22.0	2.5–4.4	3.45
6– 8	7.0		
9–11	10.0		

QUESTIONS AND PROBLEMS

1. Define the following concepts:

Array
Quantitative Grouping
Frequency Distribution
Class Interval
Class Width
Rounded Limits
True Limits
Class Midpoint
Grouping Error
Tabulation
Stub
Caption
Body
Per cent Frequency
Cumulative Frequency Distribution
Cumulative Frequency
"Less Than" Cumulation
"Or More" Cumulation

2. (a) State considerations governing the choice of the class interval.
(b) Why is it good practice to make class intervals equal in size?

3. (a) When are unequal class intervals appropriate? Give several examples.
(b) State considerations governing the construction of unequal intervals.

4. (a) When are small class intervals appropriate?
(b) Illustrate and discuss large class intervals.
(c) Are class intervals intrinsically small or large?

5. Must the lowest class interval in every tabulation begin with zero?

6. Distinguish between rounded and true class limits.

7. Is the distinction between rounded and true class limits applicable to discrete data? Explain.

8. Give two illustrations of groupings "adapted to the nature of the data."

9. How do you determine whether there are too many (few) intervals?

10 (a) How do you recognize whether class limits are true or rounded?
(b) From the rounded limits is it possible to determine the manner of rounding? Explain.

11. The delinquency rates listed below are for 140 local areas in Chicago (Table 3.1.7). Each variate is the number of boys per 100 making a court appearance.
(a) Round to the nearest whole.
(b) Construct a frequency table: rounded lower limit, 0; class width, 2.
(c) Write the true class limits.
(d) Convert class frequencies to per cents, which should sum to 100.

16. Construct an "or more" cumulative frequency distribution from Table 3.1.9.

Table 3.1.9

Distribution of Families by Size, U.S., 1950

SIZE OF FAMILY	PER CENT
2 persons	32.5
3 "	22.8
4 "	20.9
5 "	12.3
6 "	5.9
7 or more	5.6
TOTAL	100.0

Source: *Current Population Reports, Population Characteristics*, Series P-20, No. 75, U.S. Government Printing Office, Washington, D.C., June 9, 1957.

17. (a) Express the number of Indians in each age group (Table 3.1.10) as a percentage of the total.
    (b) Combine percentage frequencies as necessary to compare the age distribution of Indians in Table 3.1.10 with the 1950 United States age distribution of Table 3.1.5.

Table 3.1.10

Age Distribution, American Indians, U.S., 1950

AGE	NUMBER (IN THOUSANDS)
0– 4	52
5– 9	44
10–14	44
15–19	34
20–24	30
25–29	23
30–34	20
35–39	19
40–44	15
45–49	14
50–54	12
55–59	9
60–64	8
65–69	7
70–74	5
75 and over	6
TOTAL	342

Source: U.S. Bureau of the Census, *U.S. Census of the Population: 1950*, Vol. IV, *Special Reports*, Part 3, Chapter B, Nonwhite Population by Race, Table 3, U.S. Government Printing Office, Washington, D.C., 1953, p. 17.

14. (a) Rearrange the following class intervals and their corresponding frequencies from high to low.
    (b) Is the meaning of the frequency distribution in any way changed? Explain.

Class Limits	*Frequency (f)*
4– 6	3
7– 9	9
10–12	12
13–15	5

15. (a) Group the variates of Table 3.1.8 in class intervals of width .020 (20 points).
    (b) Describe in your own words the pattern of frequency distribution.
    (c) Compare American and National League distributions.

Table 3.1.8 *Major League Batting Averages, 1957* *

National League				American League			
.403	.285	.266	.237	.388	.276	.259	.241
.351	.284	.265	.237	.365	.276	.259	.239
.333	.284	.264	.236	.322	.276	.259	.238
.322	.283	.263	.235	.318	.275	.259	.237
.322	.283	.261	.232	.317	.274	.257	.236
.318	.283	.260	.230	.309	.273	.257	.235
.315	.279	.259	.229	.304	.272	.257	.234
.313	.279	.259	.227	.303	.272	.256	.233
.309	.278	.258	.224	.301	.270	.256	.230
.307	.277	.257	.219	.297	.270	.256	.228
.305	.277	.256	.218	.297	.270	.255	.226
.299	.275	.255	.216	.295	.270	.254	.225
.298	.274	.253	.207	.294	.269	.254	.225
.297	.274	.253	.205	.292	.268	.253	.218
.295	.273	.253	.198	.292	.268	.252	.216
.293	.273	.251	.190	.291	.268	.252	.213
.293	.273	.250	.181	.289	.267	.251	.212
.293	.272	.250	.168	.288	.267	.251	.205
.292	.272	.249	.160	.286	.267	.251	.204
.292	.271	.248		.284	.266	.251	.201
.290	.271	.247		.282	.263	.250	.194
.290	.270	.244		.281	.262	.247	.191
.289	.270	.243		.281	.262	.245	.191
.289	.268	.242		.278	.261	.245	.184
.287	.268	.240		.278	.261	.242	.180
.286	.268	.240		.277	.261	.241	

* Based on 125 or more times at bat.

Source: *New York Times*, October 2, 1957, p. 37.

tribution of the variable, analogous to the frequency distribution of the variates of the quantitative variable.

An example is provided in Table 3.2.1, where the population of the United States is grouped according to the qualitative variable of religion, whose attributes are the respective church denominations.

Compared to the quantitative tabulation, this type may appear relatively uncomplicated. Absent are class intervals, midpoints, rounded and true limits; there is no dichotomy of continuous and discrete data; there is no natural, inevitable order or array of the attributes, and therefore they cannot be cumulated. It is merely an enumeration of the frequencies of the respective attributes. But in spite of this apparent simplicity, these tabular classifications are also bound by certain rules that are designed to carry out the purposes of all classification. These purposes are: to organize a body of facts as completely as necessary and as unambiguously as possible; to set forth clearly the attributes and their relative frequencies, or *weights*. To achieve these ends fully, it must be (1) possible to classify every item, and (2) impossible to classify an item in more than one category.

Principles of Classification. The first rule, sometimes labeled the *principle of inclusiveness*, requires that the classification must contain all categories necessary to accommodate every item. It must exhaust the data so that the frequencies add up to 100 per cent. Any omission of items would have the effect of distorting the relative weights of the respective attributes, thereby producing misleading impressions. Hence, if for any reason some items remain unclassifiable — their identity may be unknown, or perhaps, in the interest of brevity, not required — they must still be accounted for in residual categories of "unknown" or "all other," so that the ratios among subgroup frequencies are preserved. This procedure is illustrated in Table 3.2.1.

The second rule demands that all items be subsumed under a single well-defined criterion, which is the subject of the table. Otherwise expressed, all items must fit one variable exclusively, and each item must fit only one attribute. When these conditions are satisfied, it will be impossible to classify an item in more than one category. This is known as the *principle of exclusiveness.* But strict compliance with this principle will be impossible when (a) two or more variables are improperly mixed within a "single" classification, or when (b) categories overlap one another — that is, when they are not mutually exclusive.

Common Violations of Principles: (1) *Mixed Variables.* The classification of "races" as *White, Negro, Indian, Mexican,* and *Jewish,* occasionally met with in less rigorous descriptions of American population, is defective in that the categories do not possess a common denominator. There are

18. Construct an "or more" cumulative distribution for the absolute frequencies given in Table 3.1.11.

Table 3.1.11

Distribution of Families by Income, U.S., 1955

FAMILY INCOME	NUMBER (IN THOUSANDS)
Under $1,000	3,300
$ 1,000–$ 1,999	4,200
2,000– 2,999	4,700
3,000– 3,999	6,300
4,000– 4,999	6,600
5,000– 5,999	5,400
6,000– 6,999	4,100
7,000– 9,999	5,500
10,000– 14,999	2,100
15,000 and over	600
TOTAL	42,800

Source U.S. Bureau of the Census, *Current Population Reports, Consumer Income*, Series P–60, No. 24, U.S. Government Printing Office, Washington, D.C., April, 1957.

19. Review the computation of an arithmetic rate (Table 3.1.1a). Express the following absolute counts as rates·
    (a) Community population, 28,362, number of births, 672. Calculate the birth rate per 1,000 general population.
    (b) Community population, 600,242; number of suicides, 50. Calculate the suicide rate per 100,000 general population.
    (c) 17,252 marriages; 4,099 divorces. Calculate the number of divorces per 100 marriages.

20. Write the true limits and class midpoints for Table 3.1.3 (distribution of blocks by number of dwelling units, Los Angeles).

SECTION TWO

Qualitative Variables

If a given descriptive term cannot be graded on a scale of more or less, higher or lower, or smaller or larger, we designate it as qualitative, to distinguish it from the quantitative categories set forth in the previous section. Thus, sex is classifiable not in terms of more or less, but rather as answers to the questions· "which one?" or "what kind?" Other qualitative variables, e.g., nationality, religious affiliation, marital status, crime, family and occupation, are also classifiable into their several attributes. The frequencies of the respective attributes constitute the frequency disresponses reducible to several variables. As Zeisel notes, "some answers refer to the qualities of the cream, some to how the respondent became acquainted with the cream, and some indicate the respondent's special needs." In general, unless tbe attributes are reducible to a common denominator, it is impossihle to maintain their mutual exclusiveness, and the comparison of their weights will be, to that extent, vitiated. A more appropriate procedure would be to tahulate separately each variable by its attributes, and therehy display the relative importance of reasons of the same essential kind. Thus, it is meaningful to compare the weights of alternative sources of information (Tahle 3.2.2b), but not to compare quantitatively a given source of information with a quality of the cream.

Table 3.2.2b

Responses by Source of Information, American Women

SOURCE	PER CENT
Recommendation	27%
Heard it advertised over radio. .	17
Saw it on the counter.	14
No answer in this category......	42
TOTAL...	100%

(2) *Overlapping Categories.* In the foregoing example, the confusion was produced by mixing distinct variables. However, in Table 3.2.3, the disorder stems more from overlapping categories: the tabulated reasons for "regret at failure to marry" are not mutually exclusive within the same variahle. In fact, in so far as ey are not virtually synonymous, some are already implied, or embraced, by others. The first item is actually a synthesis of many of the remaining categories; it is virtually impossihle to distinguish at all between No. 3 and Nos. 5, 10, 11, 12, and 15.

Now, the reader is almost bound to compare the percentages with one another, irrespective of the author's intentions. However, these percentages are not strictly comparable: Item I is not three times the significance of Item 7, since an unknown number of elements in Item 7 are already included in Item 1, thereby inflating the magnitude of Item 1 in comparison to Item 7. It is as though a group of individuals were given the opportunity to vote their preference, with the following results:

Apples	47%
Jonathans	25
Winesaps	18
Grimes Golden	10
	100%

No one would declare that Apples were approximately five times as popular as Grimes Golden. Because of the substantive overlap, the weights of the

Table 3.2.1 *Adult Population by Religion Reported, Sample U.S., 1957*

RELIGION	NUMBER	PER CENT
Protestant	78,952,000	66.2%
Baptist	*23,525,000*	*19.7*
Lutheran	*8,417,000*	*7.1*
Methodist	*16,676,000*	*14.0*
Presbyterian	*6,656,000*	*5.6*
Other Protestant	*23,678,000*	*19.8*
Roman Catholic	30,669,000	25.7
Jewish	3,868,000	3.2
Other religion	1,545,000	1.3
No religion	3,195,000	2.7
Religion not reported	1,104,000	0.9
TOTAL	119,333,000	100.0%

Source: U.S. Bureau of the Census, *Current Population Reports*, Series P-20, No. 79, U.S. Government Printing Office, Washington, D.C., February, 1958.

at least two dimensions which compete for the assignment of items: the physical race and the cultural nationality. Many Mexicans, and nearly all Jews, are *Caucasoids; other Mexicans* are *Indians*, or *Mongoloids*. Similarly, Ferri's famous, but now obsolete classification of criminals as *Insane, Born Criminal, Habitual, Occasional*, and *Criminal by Passion* is even more difficult to apply. Biologic, personality, and cultural variables are indiscriminately mixed in the same classification.

Another example of mixed variables is provided by Table 3.2.2a, in which reasons for buying face cream are set forth along with the percentage of women mentioning each one. While this contrived tabulation may appear to be free from faults, actually it is a collection of miscellaneous

Table 3.2.2a

Reasons Given for Buying Face Cream, 250 American Women

REASON	PER CENT
Recommendation	27%
Beneficial to skin	20
Heard it advertised over radio	17
Saw it on the counter	14
Reasonably priced	9
Scent appealed	7
Because of special skin condition	6
TOTAL	100%

Source: Hans Zeisel, *Say It With Figures*, 4th ed., Harper & Brothers, New York, 1957, p. 165 (adapted).

to know: wbo bis family is; how much money he has; what sort of education be bas; or bow he believes and feels about certain things?" Their replies were summarized as shown in Table 3.2.4.

The layout of this table gives every appearance of being a conventional frequency distribution on one specific variable. Tbe format entices tbe reader to read down tbe column and spontaneously to add the percentages, which do not sum to 100, but rather add up to almost 130. In fact, in a footnote, tbe autbor warns against obeying that impulse, since an unknown number of persons were tabulated in more than one category. However, tbat device nullifies tbe very purpose of the table: namely, to establish the relative importance of the various class criteria. The author treats tbe tabulation as proof tbat subjective beliefs, rather than objective circumstances such as money, family, and education, are tbe predominant earmark of social class. Multiple entries, however, introduce an element of confusion wbich disallows such a pat conclusion — and this for two reasons: (1) plural votes are accorded equal weigbt, whicb they almost certainly do not deserve; and (2) the uncontrolled number of entries which were permitted each respondent may reflect the imagination and loquacity of the respondents as much as their personal convictions. In the extreme case, if all the respondents had voted for all alternatives, there would have been no differentiation at all.

Table 3.2.4

Criteria for Social Class Membership, Sample of U.S. Population

Criterion	Per Cent Reporting *
Beliefs and Attitudes	47.4
Education	29.4
Family	20.1
Money	17.1
Other Answers	5.6
Don't Know	9.1

* Percentages add to more than 100%. People often gave more than one answer.

Source: Richard Centers, *The Psychology of Social Classes*, Princeton University Press, Princeton, N.J., 1949, p. 91.

To avoid such a dilemma, most investigators prudently restrict the answers to tbe "one best response," or tabulate tbe responses by preferential order. If tbis had been done, it is conceivable that the criterion of "Beliefs and Attitudes," which was mentioned by nearly half tbe group, and hence collectively takes first place, might rarely or never have been given first place by the separate individuals. Thus, the tabulation of the "one best response" could have reduced, or even wiped out, the attribute which now stands at the head of the column. Of course, if for any reason multiple responses are wanted, then it is incumbent on the researcb designer

Table 3.2.3 *Reasons for Regret at Failure to Marry, Female College Graduates, 1928*

Item No.	Reason	Number	Per Cent
1.	Belief marriage normal life for woman	167	31.4%
2.	Desire for children	89	16.7
3.	Desire for husband	19	3.6
4.	Desire for husband and children	49	9.2
5.	Desire for home	9	1.7
6	Desire for home and husband	2	.4
7	Desire for home and children	57	10.7
8.	Desire for home and family life	2	.4
9	Desire for husband, home, and children	32	6.0
10	Believes would be happier married	14	2.6
11.	Desire for physical relationship	14	2.0
12	Desire for married life	5	.9
13	Regret having missed a richer, fuller experience	28	5.3
14	Have domestic tastes	0	1.1
15.	Dread of loneliness	12	2.3
16.	Regret not having found the right man.	0	1.1
17.	Reason not given by others	21	4 0
	Total	532	100 0%

Source Katherine Bement Davis, "Why They Failed to Marry," *Harpers*, CLVI, 1928, p. 468 (adapted).

tabulated percentages do not reliably reflect the relative importance of the professed choices. In general, whenever the respondents have not been provided with a discriminating schedule to express themselves in an unambiguous way, we must provide major and minor subheads as exemplified in Table 3.2.1, which distribute the detailed characteristics as subtotals in italics.

Neither one of the last two tables is therefore a table in the model sense. They are only a listing of one-way percentages, to be read across the rows rather than vertically, since vertical comparisons are meaningless. If it is agreed that the intention of a table is to set forth relative weights, it is absolutely essential that categories fit one variable, and that they be mutually exclusive

(3) *Multiple Entries.* When we tolerate mixed variables or overlapping attributes, there will result the tempting opportunity to make more than one entry per case. This will, of course, result in a larger number of entries than there are respondents, with consequent still greater confusion in relative weights of attributes and uncertainty in interpretation.

In a study of social class, persons were asked: "In deciding whether a person belongs to your class or not, which of these things is most important to know: who his family is; how much money he has; what sort of education he has; or how he believes and feels about certain things?" Their replies were summarized as shown in Table 3.2.4.

The layout of this table gives every appearance of being a conventional frequency distribution on one specific variable. The format entices the reader to read down the column and spontaneously to add the percentages, which do not sum to 100, but rather add up to almost 130. In fact, in a footnote, the author warns against obeying that impulse, since an unknown number of persons were tabulated in more than one category. However, that device nullifies the very purpose of the table: namely, to establish the relative importance of the various class criteria. The author treats the tabulation as proof that subjective beliefs, rather than objective circumstances such as money, family, and education, are the predominant earmark of social class. Multiple entries, however, introduce an element of confusion which disallows such a pat conclusion — and this for two reasons: (1) plural votes are accorded equal weight, which they almost certainly do not deserve; and (2) the uncontrolled number of entries which were permitted each respondent may reflect the imagination and loquacity of the respondents as much as their personal convictions. In the extreme case, if all the respondents had voted for all alternatives, there would have been no differentiation at all.

Table 3.2.4

Criteria for Social Class Membership, Sample of U.S. Population

CRITERION	PER CENT REPORTING *
Beliefs and Attitudes......	47.4
Education.	29.4
Family..................	20.1
Money..................	17.1
Other Answers...........	5.6
Don't Know.............	9.1

* Percentages add to more than 100%. People often gave more than one answer.

Source: Richard Centers, *The Psychology of Social Classes*, Princeton University Press, Princeton, N.J., 1949, p. 91.

To avoid such a dilemma, most investigators prudently restrict the answers to the "one best response," or tabulate the responses by preferential order. If this had been done, it is conceivable that the criterion of "Beliefs and Attitudes," which was mentioned by nearly half the group, and hence collectively takes first place, might rarely or never have been given first place by the separate individuals. Thus, the tabulation of the "one best response" could have reduced, or even wiped out, the attribute which now stands at the head of the column. Of course, if for any reason multiple responses are wanted, then it is incumbent on the research designer to identify and tabulate them accordingly so that clear communication is assured

The Concept of a Table. It is clear that there is more than mere "shape" to a table, and that its construction requires more than routine skill in layout and spatial arrangement. A table is much more than a mere listing; it is an organization of the data. A well-built table constitutes the pith and meaning of a study in its definitive form. Consequently, serious consideration of its form and content should not be postponed to the final stages of a study, when the tabular report becomes due. To assure advance consideration of tabular requirements, the preparation of dummy tables is recommended. We must reconcile ourselves to the fact that the ultimate reader may lift the table out of context for purposes of analysis and quotation; it must therefore be as nearly self-contained and autonomous as possible

Adoption of principles of sound table construction, furthermore, will enforce sound conceptual practice. The anticipation of rigorous tabulation will promote rigorous research design; on the other hand, habits of flabby table construction will obstruct critical and careful planning.

It is the qualitative variable that is particularly susceptible to equivocal treatment and interpretation, since concepts used in the investigation of attitudes, motives, and preferences, for example, are not standardized and usually have their origins in common sense. Such is typically not the case with concepts covering quantitative phenomena like income and age, which are automatically reducible to mutually exclusive variates. In an age distribution, it is impossible for a person to occupy two places at the same time in the array.

As for the reader of the table, he too should approach the table critically. A useful technique is to seek out the conspicuous item, study the relative magnitudes or frequencies, and analyze the elements in the title in relation to the content of the table. In this manner, the essentials of the table that are relevant to the purposes of the reader are easily extracted.

QUESTIONS AND PROBLEMS

1 Define the following concepts:
   - Qualitative Grouping
   - Principle of Inclusiveness
   - Principle of Exclusiveness
   - Mixed Variables
   - Overlapping Categories
   - Multiple Entries

2. (a) Why is it reasonable to list attributes according to magnitude of frequency?
   (b) Cite instances when this rule may not be observed.

3. Criticize the following mock table:

Most Important Reason Given by Students for Attending College	*Per Cent*
Desire for education	30%
Cultural development	24
Preparation for job	20
Social status	18
Find a mate	8
Total	100%

4. (a) List the salient features of the tabulation of employed persons by reason for not working (Table 3.2.5).
   (b) Explain the given order of attributes.
   (c) How can the distribution of percentages be accounted for?

Table 3.2.5

Reasons for Not Working, Employed Persons, U.S., July 7–13, 1957

Reason	Number ('000)	Per Cent
Bad weather	17	0.2%
Industrial dispute	113	1.6
Vacation	5,577	79.6
Illness	793	11.3
All other	514	7.3
Total with job but not at work	7,014	100.0%

Source. U.S. Bureau of the Census, *Current Population Reports*, Series P-57, No. 181, U.S. Government Printing Office, Washington, D.C., August, 1957.

5. Evaluate Table 3.2.6 (on page 58) from the standpoint of
   (a) clarity of title
   (b) overlapping categories
   (c) possible use of percentages
   (d) conspicuous items
   (e) order of entries

Section Three

Cross-Classification

Function of Cross-Classification. Social analysis is rarely limited to the portrayal of a single variable in successive class intervals or categories but usually extends to the *association* between two or more variables. For

Table 3.2.6 *Most Important Reason for Having Last Child, Husbands and Wives, Indianapolis, 1950*

Reason Rated Most Important	Wives	Husbands
A strong liking for children	397	503
A belief that it is a religious duty to have a family	30	47
The traditional belief that married couples ought to have children	123	124
A feeling that it is important to carry on the family name	8	29
A desire to see what own children would be like	68	47
A feeling that children bring husband and wife closer together	317	241
Not wanting an only child	167	131
Not to be left childless in case of death of only child	14	5
The desire of children for more brothers and sisters	32	21
Wanting a girl if only had boys, or a boy if only had girls	71	75
Unknown	27	41
Totals	1,254	1,257

Source: Ronald Freedman and P. K. Whelpton, "Social and Psychological Factors Affecting Fertility," *Milbank Quarterly*, July 1950, p. 430 (adapted).

such purposes *bivariate*, or even *multivariate*, classifications are required *in place of the simpler univariate tabulation. Cross-classification*, which is here broached, permeates all statistical analysis; hence, it will be frequently encountered in one guise or another throughout this text.

The distribution of suicide rates according to magnitude supplies an important breakdown of the data, but it also raises new questions such as: why are the rates in some cities higher than in others? why do the rates mass around 12 per 100,000 and thin out on the extremes? These questions assume the existence of factors that determine the differential magnitudes of the rates, and challenge the statistical investigator to uncover them. For purposes of unveiling such relationships, the statistical method offers a wide variety of techniques of varying complexity, one of the simplest of which is the *cross-tabulation*. This procedure simultaneously groups a set of items by two or more criteria instead of only one. The resulting frequencies — called *joint frequencies* because they record the number of joint occurrences — give clues to how the variables may be linked, and they may often be more effectively employed than more advanced, complex correlational techniques.

Types of Cross-Classification and Their Interpretation. In Table 3.3.1a, the 107 cities in Table 3.1.1a have now been cross-tabulated by size of suicide rate and geographical region on the plausible supposition that regional differences may at least partially account for the variations observed in the univariate distribution. It is evident from the percentage distributions (Table 3.3.1b) that the cities of the Northeast, South, and North Central regions do not differ markedly from one another. However, the cities of the West exhibit a very deviant pattern: all fifteen of the western cities, with one exception (in the interval 12–14), are in the upper half of the distribution. They are conspicuously higher than the cities of the other regions. Apparently, there are social and demographic conditions in the western cities that increase the incidence of suicide. We would logically turn to the characteristics of these cities in the effort to account for the variation in rates. The cross-classification has served its preliminary purpose.

The foregoing table was made up of one quantitative and one qualitative variable; but two qualitative variables may be cross-tabulated in the same way. Thus, marriages may be cross-classified by religion of husband and of wife in order to ascertain the extent of marriages within the faith (Table 3.3.2a). In the illustrated sample of 437 marriages, 398 are of a common faith. The relative degree of religious homogamy is better expressed in Table 3.3.2b, where cell frequencies have been converted to percentages of the grand total. Summing the diagonal percentages, left to right, we obtain 92 per cent — which neatly summarizes the extent of religious endogamy in this sample.

Table 3.3.1a *Suicide Rates by Magnitude and Region, 107 American Cities*

Suicide Rate	Region				Total
	N. East	*South*	*N. Central*	*West*	
3– 5	4	2	0	0	6
6– 8	6	9	3	0	18
9–11	11	9	9	0	29
12–14	8	7	8	1	24
15–17	1	3	7	2	13
18–20	2	1	0	4	7
21–23	0	0	1	3	4
24–26	0	1	0	3	4
27–29	0	0	0	2	2
Total	32	32	28	15	107

Source: U.S. Bureau of the Census, *U.S. Census of the Population: 1950*, Vol. II, General Characteristics, Part I, U.S. Government Printing Office, Washington, D.C., 1953.

But the apparent simplicity of a percentage statement should not be permitted to obscure its hazards. When the base total (N) is not given,

Table 3.2.6 *Most Important Reason for Having Last Child, Husbands and Wives, Indianapolis, 1950*

Reason Rated Most Important	Wives	Husbands
A strong liking for children	667	593
A belief that it is a religious duty to have a family.. .	30	47
The traditional belief that married couples ought to have children	123	124
A feeling that it is important to carry on the family name..	8	29
A desire to see what own children would be like	68	47
A feeling that children bring husband and wife closer together	147	244
Not wanting an only child	167	131
Not to be left childless in case of death of only child ...	14	5
The desire of children for more brothers and sisters.... .	32	21
Wanting a girl if only had boys, or a boy if only had girls. .	71	75
Unknown	27	41
Totals	1,354	1,357

Source Ronald Freedman and P K Whelpton, "Social and Psychological Factors Affecting Fertility," *Milbank Quarterly*, July 1950, p. 430 (adapted).

such purposes *bivariate*, or even *multivariate*, classifications are required in place of the simpler *univariate* tabulation. Cross-classification, which is here broached, permeates all statistical analysis; hence, it will be frequently encountered in one guise or another throughout this text.

The distribution of suicide rates according to magnitude supplies an important breakdown of the data, but it also raises new questions such as: why are the rates in some cities higher than in others? why do the rates mass around 12 per 100,000 and thin out on the extremes? These questions assume the existence of factors that determine the differential magnitudes of the rates, and challenge the statistical investigator to uncover them. For purposes of unveiling sucb relationships, the atatistical method offers a wide variety of techniques of varying complexity, one of the simplest of which is the cross-tabulation. This procedure simultaneously groups a set of items by two or more criteria instead of only one. The resulting frequencies — called *joint frequencies* because they record the number of joint occurrences — give clues to how the variables may be linked, and they may often be more effectively employed than more advanced, complex correlational techniques.

Types of Cross-Classification and Their Interpretation. In Table 3.3.1a, the 107 cities in Table 3.1.1a have now been cross-tabulated by size of suicide rate and geographical region on the plausible supposition that regional we have no means of knowing whether the several percentages rest on 25, 50, or 500 cases, and to that extent therefore cannot judge their dependability. Percentages based on small absolute numbers are of doubtful reliability. For this reason it is good practice to include the total (N) with every percentage distribution, so that actual frequencies can be reconstructed. In addition, actual frequencies may be recombined, whereas percentages on different bases may not be so manipulated.

A variant of the foregoing tabulation is one that expresses cell frequencies as percentages of either row or column totals, usually referred to as *marginal totals*. We sometimes select those marginals representing the "causal," or independent, variable. For example, the avowed objective of Table 3.3.3 was to discover whether the opinion of the interviewer exerts an influence on the answer he obtains. Since in this instance the hypothetical causal factor (interviewer opinion) lies in the rows rather than the columns, frequencies are expressed as percentages of the row totals.

Table 3.3.3 *Respondent Attitude Toward Intervention, by Interviewer Attitudes (Percentage Distribution)*

INTERVIEWER ATTITUDE	RESPONDENT ATTITUDE		TOTAL
	"Keep Out"	"Help England"	
"Keep Out"	56	44	100%
"Help England"	40	60	100%

Source. Hadley Cantril, *Gauging Public Opinion*, Princeton University Press, Princeton, N.J., 1944, p 107 (adapted).

The data reveal that, to the "Keep Out" interviewers, a majority (56 per cent) reply similarly with the "Keep Out" slogan, whereas the "Help England" interviewers elicit a majority (60 per cent) of "Help England" responses. Apparently, the opinion of the interviewer exerts some influence on the response distribution, either because his opinion is unconsciously contagious, or because he unwittingly misclassifies the ill-defined and ambiguous responses in accordance with his own attitudes.

An example of the cross-classification of two quantitative variables is presented in Table 3.3.4, in which 1,848 marriages are distributed by age of husband and by age of wife. In keeping with the purposes of cross-classification, we examine the joint frequencies in order to determine whether they form a discernible pattern. There is a fairly obvious trend: husbands tend to be slightly older than their wives. For example, of the 719 husbands ages 20–24, 153 (21.3 per cent) are married to younger women, 62 (8.6 percent) to older; of the 850 wives ages 20–24, only 10 (1.2 per cent) are married to younger men but 336 (39.5 per cent) to older.

Table 3.3.1b *Suicide Rates by Magnitude and Region, Percentage Distribution*

SUICIDE RATE	REGION				TOTAL
	N. East	*South*	*N. Central*	*West*	
3– 5	12.5%	6.3%	0.0%	0.0%	5.6%
6– 8	18.8	28.1	10.7	0.0	16.8
9–11	34.4	28.1	32.1	0.0	27.1
12–14	25.0	21.9	28.6	6.7	22.4
15–17	3.1	9.4	25.0	13.3	12.2
18–20	6.2	3.1	0.0	26.7	6.6
21–23	0.0	0.0	3.6	20.0	3.7
24–26	0.0	3.1	0.0	20.0	3.7
27–29	0.0	0.0	0.0	13.3	1.9
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%

Source See Table 3.3.1a.

Table 3.3.2a *Marriages by Religion of Husband and Wife, 437 Married Couples, New Haven, Conn.*

HUSBAND	WIFE			TOTAL
	Catholic	Protestant	Jewish	
Catholic	271	20	0	291
Protestant	17	61	0	78
Jewish	1	1	66	68
TOTAL	289	82	66	437

Source A. B. Hollingshead, "Cultural Factors in the Selection of Marriage Mates," *American Sociological Review*, XV, 1950, p. 623.

Table 3.3.2b *Marriages by Religion of Husband and Wife (Percentage Distribution)*

HUSBAND	WIFE			TOTAL
	Catholic	Protestant	Jewish	
Catholic	62	4	0	66
Protestant	4	15	0	19
Jewish	0	0	15	15
		19	15	100%

(b) Does there appear to be a relationship between sex and occupation? Explain.

Table 3.3.5 *Major Occupation Group of Employed Persons by Sex, U.S., July 7–13, 1957*

Major Occupation Group	Number *			Per Cent		
	Male	*Female*	*Both Sexes*	*Male*	*Female*	*Both Sexes*
Professional, technical, and kindred workers	3,964	2,165	6,129	8.7%	10.1%	9.1%
Farmers and farm managers	3,329	195	3,524	7.3	0.9	5.2
Managers, officials, and proprietors, except farm . .	5,912	1,001	6,913	12.9	4.7	10.3
Clerical and kindred workers	3,022	6,280	9,302	6.6	29.2	13.8
Sales workers	2,543	1,616	4,159	5.6	7.5	6.2
Craftsmen, foremen, and kindred workers	8,543	256	8,799	18.7	1.2	13.1
Operatives and kindred workers	9,156	3,476	12,632	20.0	16.2	18.8
Private household workers	50	1,971	2,021	0.1	9.2	3.0
Service workers, except private household	2,832	2,883	5,715	6.2	13.4	8.5
Farm laborers and foremen	2,381	1,582	3,963	5.2	7.4	5.9
Laborers, except farm and mine.	3,981	83	4,064	8.7	0.4	6.0
Total Employed	45,713	21,508	67,221	100.0%	100.0%	100.0%

* Thousands of persons 14 years of age and over.

Source: U.S. Bureau of the Census, *Current Population Reports*, Series P-57, No. 181, U.S. Government Printing Office, Washington, D.C., August, 1957

Table 3.3.6 *Type of Family Organization by Type of Economy, 194 Primitive Societies*

Type of Economy	Type of Family			Total
	Maternal	Paternal	Intermixed	
Hunting	30	18	22	70
Pastoral	1	10	3	14
Agricultural	44	47	19	110
Total	75	75	44	194

Source: L. T. Hobhouse, G. C. Wheeler, and M. Ginsberg, *The Material Culture and Social Institutions of the Simpler Peoples*, Chapman and Hall, London, 1915, p. 153.

Other entries may be similarly examined for their possible sociological meaning

Table 3.3.4 *Age at Marriage, Husband and Wife, New Haven, Conn.*

Age of Husband	Age of Wife								Total
	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50+	
15–19	42	10	3						55
20–24	153	504	51	10	1				719
25–29	52	271	184	22	7	2			538
30–34	5	52	87	69	13	5			231
35–39	1	12	27	29	21	2	3		95
40–44		1	9	18	17	8	2	1	56
45–49	1		3	6	16	16	7	1	49
50 & Over			1	4	11	15	21	43	95
Total	254	850	365	168	86	47	33	45	1,838

Source: A. B Hollingshead, *op. cit.*, p. 622

The construction of a joint frequency table, such as those illustrated above, incorporates no new principles; it is guided by the same considerations that govern the univariate frequency table. The number and the size of the class intervals should be fixed with a view to attaining as much simplicity as possible, without at the same time distorting or blurring the relationship between the two variables. With qualitative data, the categories should be exhaustive and mutually exclusive. If such simple principles are used as guides, the preparation of a cross-tabulation should present no difficulties, and the final result should be both competent and effective.

Questions and Problems

1. Define the following concepts:
   Cross-Classification
   Cross-Tabulation
   Univariate Distribution
   Bivariate Distribution
   Multivariate Distribution
   Joint Frequency
   Joint Frequency Distribution
   Independent Variable
   Dependent Variable

2. Study Table 3.3.5.
   (a) List the important and conspicuous entries.

University Press, Princeton, N.J., 1949, p. 589.) Arrange these results in a 4 × 2 table and attach an appropriate title.

7. Cross-tabulate the paired variates of Table 3.3.8, using class intervals of $5.00 and one-half (.5) school year, and lower rounded limits of $15 and 7.5 school years respectively. Each pair of values characterizes a given census tract.

Table 3.3.8 *Median Years of School Completed by Median Monthly Rentals, 115 Census Tracts, Indianapolis, 1950*

School	Rental	School	Rental	School	Rental	School	Rental
8.7	$29	12.4	$52	9.5	$33	8.1	$32
9.2	33	12.7	61	8.5	24	9.0	29
9.6	30	12.6	61	8.7	27	8.7	32
12.1	44	12.7	62	10.3	38	8.7	30
11.3	37	12.5	46	8.5	28	8.7	27
8.1	22	12.3	68	12.5	51	8.5	24
8.4	24	12.6	57	11.5	43	9.4	32
0.1	27	12.4	61	10.8	36	9.9	32
10.7	36	12.6	55	10.2	36	9.0	28
8.8	24	14.6	74	12.2	46	10.1	37
8.8	18	12.8	60	12.1	50	12.4	43
8.5	21	12.8	57	10.7	29	12.4	49
11.7	40	12.6	58	8.4	18	12.7	87
11.3	47	12.4	57	9.0	27	12.9	92
9.0	40	12.1	40	8.4	24	12.6	51
11.9	47	12.0	37	8.5	30	12.1	35
12.3	48	12.0	37	8.3	22	8.8	30
8.4	22	10.7	36	8.2	21	11.8	35
11.5	47	10.0	35	9.4	28	10.7	37
8.5	21	9.8	38	9.1	33	9.8	36
10.4	32	8.9	34	8.9	28	9.0	29
10.4	34	8.6	29	8.8	35	9.5	26
8.8	24	8.8	36	8.7	33	9.8	31
8.8	30	10.1	39	8.9	33	10.3	31
10.5	39	7.9	21	8.6	25	12.2	46
12.0	41	8.5	25	8.3	24	11.0	39
9.7	29	9.0	27	8.4	21	12.5	36
12.4	60	11.1	42	8.5	35	11.0	36
12.6	49	9.3	32	8.8	28		

Source: U.S. Bureau of the Census, *U.S. Census of the Population: 1950, Indianapolis, Indiana, Census Tracts*, P-D25, Tables 1 and 2, U.S. Government Printing Office, 1952.

8. Older people take more interest in political elections than younger persons, as suggested by Table 3.3.9. The difference is especially marked among those

3. (a) In Table 3.3.6, which variable — family organization or basic economy — would you consider independent?
   (b) Express cell frequencies as percentages of row totals. Does this enhance the effectiveness of the table? Explain.

4. (a) In Table 3.3.7, which variable would you consider independent?
   (b) How would you explain the small number of "Unhappy" ratings by both husbands and wives?
   (c) What percentage of husbands and wives agree in their ratings?

Table 3.3.7 *Marital Happiness Ratings of Husband and Wife, 252 Married Couples*

Wife's Rating	Husband's Rating					Total	Per Cent
	Very Unhappy	Unhappy	Average	Happy	Very Happy		
Very Happy	1		3	24	112	140	55.6%
Happy			12	38	12	62	24.6
Average		3	14	7	6	30	11.9
Unhappy	1	11	2			14	5.5
Very Unhappy	5	1				6	2.4
Total	7	15	31	69	130	252	
Per Cent	2.8%	5.9	12.3	27.4	51.6		100.0%

Source Ernest W. Burgess and Leonard S. Cottrell, "The Prediction of Adjustment in Marriage," *American Sociological Review*, I, 1936, p. 741.

5. A club has 100 members. Among them are 50 lawyers and 50 liars. The number of members who are neither lawyers nor liars is 20. Construct a 2 × 2 table to show the number of lawyers who are liars, the number of lawyers who are not liars, the number of liars who are not lawyers, and the number who are neither.

6. During World War II, white commissioned and non-commissioned officers were asked to compare the effectiveness of colored troops with that of white troops. The classification of answers was as follows:
   (a) Not as good as white
   (b) Same as white
   (c) Better than white
   (d) No answer

   Of the commissioned officers ($N = 60$), 5 per cent gave the first reply listed, 69 per cent the second, 17 per cent the third, and 9 per cent the fourth. Of the noncommissioned officers ($N = 195$), the distribution was 4 per cent, 83 per cent, 9 per cent, and 4 per cent. (Source: Samuel A. Stouffer *et al.*, *The American Soldier*, Vol. 1 of *Studies in Social Psychology in World War II*, Princeton of social change, are as significant to the sociologist as is the description of its static aspects. Not only are we interested in the momentary size of the American population, but also in its growth or decline. The social analyst does not restrict his attention to the divorce rate in a given year, but necessarily also examines its fluctuations from year to year. Social trends are as revealing of the laws of social behavior as are the cross-sections of social life.

The time series (Table 3.4.1) is one of the simpler devices for the portrayal of cultural and social changes; it is a set of values systematically arranged in chronological order, and is thereby distinguished from a mere *chronicle of events. Nevertheless, the construction of a time series is* governed by technical considerations analogous to those entering into the preparation of the frequency distribution: (1) the intervals should be of such length as to depict the essential trend or pattern without needless detail; (2) they should be expressed in familiar units of time and their multiples, such as the month or year; and (3) they should be uniform in size and, when possible and fitting, evenly spaced throughout the series. Furthermore, the variable should have a constant definition for the entire period covered by the series, and the boundaries of the geographic area in which the events occur should also remain fixed throughout. These rules will be more difficult to enforce as the time series increases in duration, owing to the malleability of social circumstance.

Table 3.4.1

Number of Divorces per 100 Marriages, U.S. Selected Years, 1870–1955

Year	Rate
1870	3.1
1880	4.3
1890	5.8
1900	7.9
1910	8.8
1920	13.4
1930	17.4
1940	16.5
1945	30.8
1950	23.1
1955	24.6

Source: U.S. Department of Health, Education, and Welfare, *Vital Statistics of the United States, 1950* and *1955*, Vol. I, U.S. Government Printing Office, Washington, D.C.

The Time Interval. There is, of course, no standard time interval for all occasions; rather the interval will vary according to the requirements of the investigation. Thus, the purpose may be to determine the hourly having more education, but it is also present among those with less education.

(a) Condense Table 3.3.9 to show the percentage distribution of interest in elections by education, irrespective of age. (Suggestion: convert given percentages into absolute frequencies, recombine as necessary, and convert to percentages)

(b) What information is "lost" in the condensation?

Table 3.3.9 *Interest in Political Elections, by Education and Age*

Degree of Interest	Education			
	No High School		Some High School or More	
	Under 45 Yrs.	*45 or Older*	*Under 45 Yrs.*	*45 or Older*
Great	19%	25%	26%	41%
Little	81	75	74	59
Total	100%	100%	100%	100%
N	376	869	1,174	439

Source P Lazarsfeld, B Berelson, and H. Gaudet, *The People's Choice*, Columbia University Press, New York, 1948, p 44

9. In Table 3.3.2a, what per cent
   (a) of Catholic husbands marry Catholic wives?
   (b) of Catholic wives marry Catholic husbands?
   (c) of Protestant husbands marry Catholic wives?
   (d) of Catholic husbands marry Protestant wives?
   What light do these figures throw on the tendency of men and women to cross the religious curtain?

10. (a) Of husbands 25-29, calculate the percentage married to younger, to older, and to women of the same age (Table 3.3.4).
    (b) Of wives 25-29, calculate the percentage married to younger, to older, and to men of the same age.
    (c) Comment on social significance.

Section Four

Time Series

The Nature of the Time Series. Up to this point in the discussion, we have considered the variable only at a given instant and have ignored its presumable changes in magnitude in successive intervals of time. However, the dynamics of social organization, which manifest themselves in patterns but the changing content of the variable as well. For example, a shift in the legal definition of juvenile delinquency to include seventeen-year-old boys, who had previously been excluded, would produce an apparent increase in delinquents, and create a false impression of moral decay. Needless to state, a careful worker would endeavor to adjust his data to compensate for such revisions in definition.

Similarly, the effects of improvements in diagnostic procedures on statistical fluctuations in disease rates may be confused with the effects of the increasing average age of the population, unless appropriate corrections are made. This is well illustrated in the field of cancer, where classificatory changes make it difficult to discern whether the trends in the cancer rate are real or apparent (Table 4.4.1). The prescription of a fixed variable is especially difficult to follow in the social realm, since the definitions of many of our social variables are undergoing almost constant revision. The definition of delinquency is much less stable than that of a chemical.

Constancy of Area. It is equally desirable that the time series be based on a relatively fixed geographical area; otherwise the sequence of measures may merely reflect alterations in area boundaries. Twentieth century demographers have actually encountered such difficulties in measuring the population growth of European nations because their boundaries have been so frequently and drastically redrawn.

Similarly, between 1915 and 1933 in the United States, there was considerable unreliability in the calculated trend of the national birth rate because of the periodic addition of states to the birth registration area (Table 3.4.3). In 1915 the rate was based on the returns from only ten states; not until 1933 were all states included. Subsequently, national rates have been prepared, incorporating natality estimates for the nonregistration states, an adjustment made necessary by the lack of uniformity in the registration area. All yearly rates are normed to the same base area; hence, we call this a *norming operation*.

Measures of Change: Relative and Absolute. A time series is a succession of chronologically spaced observations, and is designed to depict growth, decline, or simply variations in the incidence of the things observed. These quantitative observations may be in the form of absolute values or relative values. For example, the rise in divorce may be measured as absolute increase in the number of divorces, or as a rise in the rate of divorces per 100 marriages. These changes are not necessarily in the same direction; nor do they even supply identically useful information on questions which the sociologist may raise. If this fact is not appreciated, the worker will often come up with inappropriate answers.

Table 3.4.4 displays the basic observations on which we may compute the growth rate in the population 65 years of age and older. There are variation in telephone calls or automobile traffic, or the daily variation in school attendance, or in industrial absenteeism. In studies of *seasonal variation* (Table 3.4 2), the month is a convenient unit, whereas the year or decade may be an appropriate interval to depict a long-term trend, such as the growth of the nation's population. Very long term trends in which fluctuations are smoothed out are sometimes known as *secular trends.*

Table 3.4.2

Number of Marriages and Marriage Rates, by Month, U.S., 1956

Month	Number	Rate *
January....... .	101,000	7.1
February.......	99,000	7.5
March.	102,000	7.2
April..........	117,000	8.5
May	128,000	9.0
June	201,000	14.7
July............	128,000	9.9
August.........	159,000	11.2
September....:.	145,000	19.6
October....	130,000	9.2
November	126,000	9.2
December......	133,000	9.4
Total.....	1,569,000	9.4

* Rates on annual basis per 1,000 midyear population, adjusted for length of month.

Source· U.S. National Office of Vital Statistics, *Vital Statistics· Special Reports, Marriages and Divorces*, 1956, Vol. 49, No. 3, Table E, U.S. Government Printing Office, Washington, D.C., 1958.

Equal intervals render successive entries comparable and intelligible. It would be impossible to establish the relative preponderance of June marriages, if all the other intervals were very irregular. In addition, the month is a sufficiently brief period to throw into relief the seasonal change, but not too extended to conceal the meaningful fluctuations. It is obvious that discriminating attention must be given the optimum size of the intervals. While the neonatal death rate of infants must be calculated on a daily basis, such refinement would be superfluous in measuring the mortality of the general population

Constant Definition of Variable. No time series will be completely adequate unless the definition of the variable remains at least approximately constant throughout the period covered by the series. Such variables as unemployment, crime, and occupation are very susceptible to shifting meanings and definitions within the course of time. In such instances, the trend in the series may mirror not only substantive changes in frequency or volume,

Table 3.4.4 *Persons 65 Years and Over, U.S. Population, Selected Years, 1900–1940*

YEAR	TOTAL POPULATION (IN THOUSANDS)	PERSONS 65+ (IN THOUSANDS)	PERCENTAGE OF TOTAL POPULATION
1900	76,094	3,100	4.1
1905	83,820	3,504	4.2
1910	92,407	3,985	4.3
1915	100,549	4,501	4.5
1920	106,466	4,929	4.6
1925	115,832	5,788	5.0
1930	123,077	6,706	5.4
1935	127,250	7,803	6.1
1940	131,954	9,031	6.8

Source: U.S. Bureau of the Census, *Current Population Reports, Population Estimates*, Series P-25, No. 114, U.S. Government Printing Office, Washington, D.C., April 27, 1955.

of old people in 1900 (3,000,000), and arrive at an increase of 200 per cent. Both of these measures are relevant when our interest is restricted to the aged population. For example, the manufacturer of hearing aids will be interested in the continued absolute growth of the older population, which constitutes for him an expanded market; he sells hearing aids to people, not to proportions. On the other hand, our interest may lie in the changing proportion of older people in the total population. National welfare administrators, for example, will want to know something about the changing ratio of dependent to productive persons. An insurance actuary will be concerned with the increasing proportion of the aged, which is indicative of a longer average length of life, for it is on this average that premiums are computed; he is more concerned with probabilities than with raw numbers. In such cases our interest shifts to Measures 3 and 4. (3) *Percentage point increase.* We see that the aged constituted 4.1 per cent of the total United States population in 1900, and 6.8 per cent of the population in 1940 — an increase of 2.7 percentage points. (4) *Per cent increase in the proportion.* To find this figure, we divide the percentage-point increase (2.7) by the percentage of old people in 1900 (4.1), and find that there has been a 66 per cent increase in the proportion of the total population 65 or over.

In certain instances, the absolute and relative measures may show opposite trends. Thus, between 1790 and 1950, the Negro in America increased from 760,000 to 15,000,000; but relative to the total United States population, he declined from 19.3 to 10 per cent. The choice of measure, therefore, must always be adapted to the substantive issue.

Table 3.4.3 *Crude Birth Rates, Registration Area and U.S., 1915–1935*

Year	Number of States in Registration Area	Birth Rate in Registration Area	U.S. Birth Rate*
1915	10	25.0	29.5
1916	11	24.9	29.1
1917	20	24.5	28.5
1918	20	24.7	28.2
1919	22	22.4	26.1
1920	23	23.7	27.7
1921	27	24.2	28.1
1922	30	22.3	26.1
1923	30	22.1	26.0
1924	33	22.2	20.1
1925	33	21.3	25.1
1926	35	20.5	24.2
1927	40	20.5	23.5
1928	44	19.7	22.2
1929	46	18.8	21.2
1930	46	18.9	21.3
1931	46	18.0	20.2
1932	47	17.4	19.5
1933	48	16.6	18.4
1934	48	17.0	19.0
1935	48	16.9	18.7

* Adjusted for underenumeration
Source U.S. National Office of Vital Statistics, *Vital Statistics of the U.S., National Summaries*, Natality Statistics, 1951, Vol. 38, No. 8, U.S. Government Printing Office, Washington, D.C., 1954, Table 1.

available to us at least four different measurements, each of which is quite legitimate for some purposes, but no single one of which can supply us with the sociologically relevant information in all situations. We may measure: (1) the absolute increase of persons 65 and over; (2) the percentage increase in the absolute growth of the population of persons 65 or over; (3) the absolute percentage point increase in the proportion of old people in the total population; and (4) the percentage increase in the proportion of old people in the total population.

Let us compare the years 1900 and 1940 in Table 3.4.4 to see what these four different measures mean. **(1)** *The absolute increase.* In 1900, there were about 3,000,000 persons 65 or older in the United States, and in 1940, there were about 9,000,000 — an absolute increase of 6,000,000. **(2)** *The percentage increase in the absolute growth.* To find the percentage increase, we divide the absolute increase of old people (6,000,000) by the number

Table 3.4.6

Divorce Rate and Percentage of Married Women Gainfully Employed, Selected Years, U.S., 1890–1950

Year	Divorce Rate	Percentage of Women Employed *
1890	5.8	4.6
1900	7.9	5.6
1910	8.8	10.7
1920	13.4	9.0
1930	17.4	11.7
1940	16.5	16.7
1945	30.8	25.6
1950	23.1	24.8

* Compiled from bulletins of the Women's Bureau, U.S. Department of Labor, U.S. Government Printing Office, Washington, D.C.

ever, that would be a premature generalization, since it would be possible that the two trends are actually a common consequence of one or more hidden factors — or it may be sheer coincidence.

In any time series of significant duration, the observed trends are always affected by innumerable, uncontrolled, and unidentified factors which make simple interpretations a precarious venture. Nevertheless, time series, when carefully constructed by keeping potentially disturbing factors constant, supply objective comprehensible measures of social change without which certain types of social analysis would be impossible.

Questions and Problems

1. Define the following concepts
   Time Series
   Seasonal Variation
   Secular Trend
   Absolute Change
   Relative Change
   Cumulated Time Series
2. Distinguish between a time series and historical chronology.
3. Cite three social variables (not given in text) subject to change of definition so that it would be hazardous to express their variation in a time series of significant duration.
4. To what extent does the correspondence between the changes in the divorce and employment rates establish a causal relation (Table 3.4.6)?
5. (a) Suggest a sociological interpretation of the seasonal variation in marriages (Table 3.4.2).
   (b) Account for the peak rate in June, the trough rate in January.
6. (a) February has fewer marriages than March but shows a higher rate (Table 3.4.2). Explain this seeming contradiction.

Cumulated Time Series. Cumulation is nothing more or less than the act of joining contiguous intervals and adding their frequencies. Under certain circumstances, the time series is subject to cumulation like a simple frequency distribution Such is the case when the separate items constitute a meaningful total frequency, distributed throughout the period, and which are cumulatively accounted for at the close of that period. An example is provided by the rate of diffusion of the postage stamp among the countries of Europe and North and South America (Table 3.4.5). Prior to 1840, only one country had adopted this device; by mid-century, nine countries were using this method; however, not until another thirty years had elapsed had all of the 37 independent countries introduced the postage stamps. Most countries waited an average length of time to adopt the postage stamp. Such a simple technique of tabulation may often be effectively utilized in a preliminary manner to reveal the course and tempo of the diffusion of cultural traits or the flow of information. In fact, such tabulations have been the occasion for the formulation of "universal" laws of growth and diffusion.

Table 3.4.5

Date of First Postage Stamp Issue, Selected Countries, 1836–1880

DATE	NUMBER	
	f	*cf*
1836–40	1	1
1841–45	2	3
1846–50	6	9
1851–55	7	16
1856–60	8	24
1861–65	6	30
1866–70	4	34
1871–75	2	36
1876–80	1	37

Source H. Earl Pemberton, "The Curve of Culture Diffusion Rate," *American Sociological Review*, I, 1936, p. 552.

Concurrent Trends Two or more time series may be juxtaposed for the purpose of exhibiting the relationship between them. This is an elementary but useful procedure which may direct preliminary attention to a possible statistical association, which may then be measured by more advanced methods.

We illustrate such a comparison by setting up the employment rate of married women alongside the divorce rate in parallel tabulation. There is a startling similarity in the rates of growth of the two series. One might therefore be tempted to conclude that the increasing divorce rate is functionally related to the increased employment of married women. How-

Graphic Presentation 4

Section One

The Histogram and the Frequency Polygon

Function of the Graph. All tabular material can be converted into graphic forms, which are usually more effective in conveying an impression of the pattern of the total distribution than is the more intricate frequency table. Being the areal equivalent of frequency, the graph gives visibility to the distribution, and consequently is more readily suggestive of its meaning and interpretation. The principal function of the graph is to convey an accurate conception of the shape of the frequency distribution — a conception which not even the skilled expert, much less the layman, could construct mentally from the raw tabular data. We cannot expect the table to do what the graph can do better.

There are many types of graphic representation, each useful in its own way when properly applied. Some of them are very ingenious and elaborate, but only the basic and simplest types will be here discussed: (1) the histogram, (2) the frequency polygon, (3) the cumulative frequency polygon, (4) the bar chart, (5) the statistical map, and (6) the time chart. Of these six types, the first three are applicable only to quantitative data. The bar chart is designed primarily for qualitative data; the statistical map displays items classified by geographical area; and the time chart is the graphic version of the time series.

Since graphs are the equivalents of tables, they should carry analogous identifying titles and markers. They are subject to the same criteria of intelligibility, simplicity, and clarity. In fact, a graph cannot be planned or constructed until the corresponding table has been prepared.

Construction of the Histogram. The histogram consists of a set of contiguous columns whose heights are proportional to class frequencies, and whose

(b) October and November have identical rates but differ in the number of marriages. Explain.

7. Is the ratio of divorces to marriages a reliable measure of incidence of marriage instability in such states as Nevada?

8. In how many ways is it possible to measure changes in volume of population under 21 years of age? Measure these changes in the U.S. population between 1900 and 1960.

9. Account for the pattern of diffusion of the postage stamp in the period 1840–1870 (Table 3.4.5).

Selected References

Jenkinson, Bruce L., *Bureau of the Census Manual of Tabular Presentation*. U.S. Government Printing Office, Washington, D C., 1949

Kemeny, John G., J. Laurie Snell, and Gerald L. Thompson, *Introduction to Finite Mathematics*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1957. Chapter 2.

Kendall, Maurice G., *The Advanced Theory of Statistics*. Volume I, fifth edition. Hafner Publishing Co., New York, 1952. Chapter 1.

Walker, Helen M., and Walter N. Durost, *Statistical Tables*. Bureau of Publications, Teachers College, Columbia University, New York, 1936.

Zeisel, Hans, *Say It with Figures* Harper & Brothers, New York, 1957. Chapters 8 and 9.

by the number of class intervals to be plotted. The result will be the number of linear units allotted to each interval. Beginning at the intersection of the two axes, we can now lay down the class intervals and mark their boundaries with tiny upright lines, or *ticks*, appropriately placed inside the axes, since they may be construed as vestigial stubs of the original grid lines. Because these ticks represent the contact points of the contiguous intervals, they are designated by markers carrying the true values of the class limits. Such true limits will necessarily be in fractional form when the values have been rounded to the nearest whole. Thus, in Figure 4.1.1, class limits are designated as 2.5, 5.5, and so forth. When, however, values have been rounded to the last whole, as in the case of United States Census age data, the true boundaries may be simply written as integers: 0, 5, 10. In the interest of clarity, markers may be uniformly spaced at selected class boundaries or placed at selected midpoints. By these discretionary procedures, the cluttered effect of crowded markers may be avoided.

The frequency scale is established on the vertical axis in an analogous manner. We teke the largest class frequency to be accomodated in the graph, and divide it by the number of linear units available on the previously drawn axis. We thereby determine the number of items to be assigned to each unit on the upright axis. Thus, in Figure 4.1.1, two items correspond to one unit. Beginning with the *zero origin* et the intersection, markers are now spaced at equal intervals, and given in such multiples (5, 10, end so on) as may be clear and intelligible.

Having marked off the two scales, we are now ready to dmw the columns of the histogram within the frame of the axes. The vertical outlines of the columns are erected on the points of the true boundaries; and their heights are determined by the frequencies of the respective class intervals.

Of the aforementioned directives, the rule that the frequency scale originate with zero is inflexible; otherwise the required proportionality among column areas could not be maintained. A violation of this rule is illustrated in Figure 4.1.2a, where the bottom half of the histogram has been amputated. The ratio of the first two class frequencies is 20 to 25, or 4 to 5. However, in Figure 4.1.2a the ratio between the areas of the first two col-

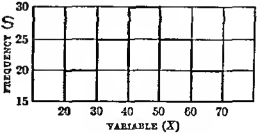


Figure 4.1.2a *Histogram (Incorrect) Without Zero Origin*

widths are proportional to the size of the class intervals of the variable. A histogram based on the frequency distribution of 107 suicide rates presented in the last chapter (Table 3.1.1) is shown in Figure 4.1.1. It is not only a graphic record of the absolute class frequencies; it also mirrors the size of each frequency relative to all others.

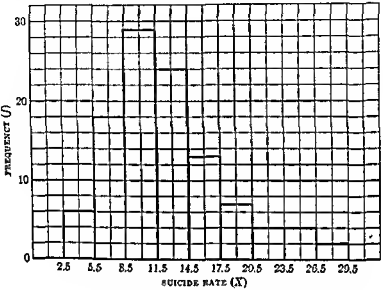


FIGURE 4.1.1 *Histogram of Suicide Rates, 107 Large U.S. Cities, 1950*

The histogram is constructed on simple *arithmetic graph paper*, which is ruled with equally spaced horizontal and vertical guide lines. These intersecting guide lines are thereby divided into equal linear units which are a necessary base for systematic plotting. The first step in the procedure is to draw at right angles on selected grid lines two axes of approximately equal length, with their intersection near the lower left-hand corner of the page Class boundaries are plotted on the *horizontal axis*, and class frequencies are plotted on the *vertical axis*.

However, before laying off the class intervals on the base line, we must fix the appropriate number of linear units to be assigned to each class interval of the table. It is often good practice to allow for a vacant interval at either end of the horizontal scale, which improves the aesthetic appearance of the graph and promotes readability. We accordingly count the total number of units along the length of the axis, and divide that total

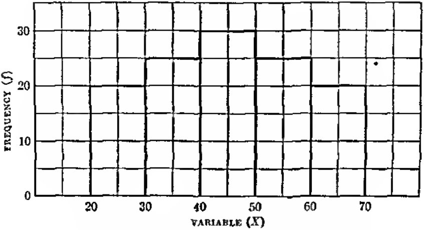


FIGURE 4.1.3a *Histogram (Incorrect), Extended Horizontal Axis*

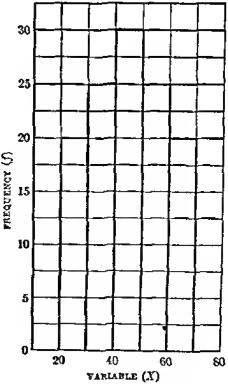


FIGURE 4.1.3b *Histogram (Incorrect), Extended Vertical Axis*

umns is 1 to 2, instead of 4 to 5, so that the areal ratio fails to correspond to the frequency ratio. The areas of the other columns in 4.1.2a are equally distorted. The correct ratios are displayed in Figure 4.1.2b.

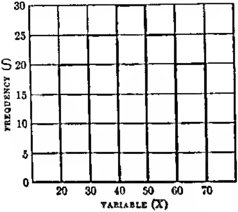


FIGURE 4.1.2b *Histogram (Correct) With Zero Origin*

However the rule that the two axes be of approximately the same length is not inflexible; it is simply considered good practice. Unless there is an obviously good reason for resorting to unequal axes, the resulting graph is likely to produce a distorted impression. By extending the base line and shortening the vertical axis, the histogram may be made to appear long and flat, thereby conveying an impression of great variation. On the other hand, by lengthening the vertical axis and contracting the base line, we produce a tall narrow figure, which leaves the impression of only slight variation. These effects are illustrated in Figures 4.1.3a and 4.1.3b, which portray the same table.

Unequal Class Intervals. In Table 4.1.1, the last interval is twice the width of the other intervals. Its frequency of four is therefore not comparable

Table 4.1.1

Frequency Distribution, Unequal Class Intervals

CLASS INTERVAL	FREQUENCY
0– 4	3
5– 9	6
10–14	14
15–19	5
20–29	4
TOTAL	32

Source: Hypothetical

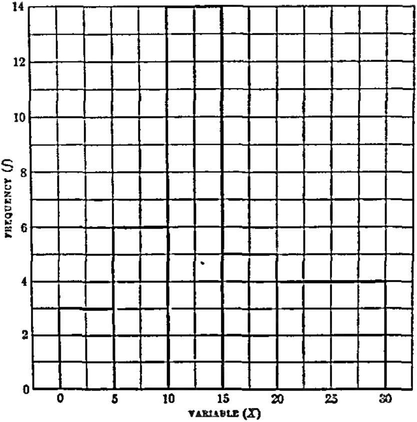


FIGURE 4.1.4b *Histogram (Incorrect), Unequal Class Intervals*

We can now appreciate the practicability of the rule set forth in Chapter 3 that class intervals be of equal width; or, failing this, that unequal class intervals be convenient multiples of smaller ones. Similarly, the caution against open-ended tables has now acquired new practical meaning. There is no way to adjust the frequency of a class interval of unknown size, since an indefinite interval cannot be expressed as a multiple of a closed interval. It is therefore impossible to represent an open-ended table by the histogram, unless one arbitrarily closes the interval — a measure to which we sometimes may resort.

The distinguishing feature of the histogram is its schematic simplicity. Columns are more easily apprehended than numbers. It clearly reveals the relative concentration of items in each interval, and shows the contour of the distribution.

to the frequencies of the smaller intervals. To establish comparability, and preserve spatial ratios, it is therefore necessary to break up the last interval into two intervals of equal width; and we must break up its frequency correspondingly into two equal frequencies of two each. These adjusted sub-frequencies are then plotted (Figure 4.1.4a). Had the unadjusted frequency of four been plotted (Figure 4.1.4b), the frequency column on the extreme right would have enclosed twice as large an area as was its due, and thereby created an impression contrary to fact.

In general, before graphing a table in which class intervals are unequal, all larger intervals must be expressed as multiples of the smallest interval; and these obtained multiples divided into the corresponding frequencies. This latter step will yield the adjusted frequencies, which are then plotted. This procedure is in accord with the principle that each item in the frequency distribution be represented by an equal area on the graph, so that relative frequencies be proportionately presented in area form.

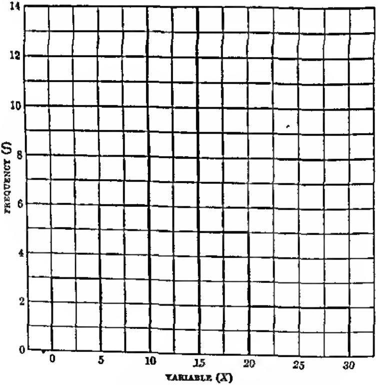


Figure 4.1.4a *Histogram (Correct), Unequal Class Intervals*

off on the base line, and the frequency scale along the vertical axis. At this stage, however, the procedure diverges, in that points and connecting lines are plotted instead of columns. The points are plotted over class *midpoints* at heights called for by the respective class frequencies, and then joined by straight lines to form the polygon.

The frequency polygon need not be extended to the base line beyond the range of actual observation. In fact, in a U-shaped distribution, in which the concentration of cases increases at either end, it would even be illogical to do so. It would appear, therefore, sound policy to leave the polygon open at either end, and thereby avoid the hazard of misrepresenting the frequencies where none were observed. In any event, the general contour of the bulk of cases is in no way affected by open ends, and it is this shape, after all, which the polygon seeks to portray.

When a closed figure is contemplated, the usual procedure is to extend the graph to the base line at the midpoints of the vacant intervals on either end (Figure 4.1.5b). The rationale of this practice is to preserve the area

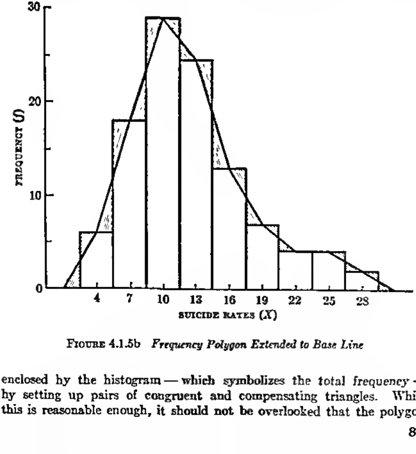


FIGURE 4.1.5b *Frequency Polygon Extended to Base Line*

enclosed by the histogram — which symbolizes the total frequency — by setting up pairs of congruent and compensating triangles. While this is reasonable enough, it should not be overlooked that the polygon

The Nature and Construction of the Frequency Polygon. When items are compressed into a relatively small number of broad intervals, the resulting class frequencies tend to jump in an abrupt manner. It is reasonable to suppose, however, that the progression of class frequencies would be considerably smoother if we employed many relatively small class intervals. It is the function of the *frequency polygon* to provide an approximation of the smooth curve that would presumably emerge if class intervals were made as small as possible and the number of observations were unlimited.

The frequency polygon may be derived from the histogram by simply connecting with straight lines the midpoints of adjacent column tops. Such a conversion is illustrated in Figure 4.1 5a, where a frequency polygon is *superimposed* on the histogram of suicide rates. In practice, only one graph or the other would be presented, depending on whether the emphasis is to be placed on the tabulated class frequencies or on the hypothetical point frequencies, which the frequency polygon provides.

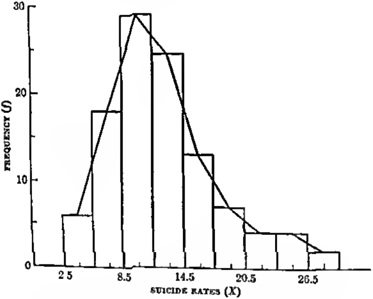


Figure 4.1.5a *Histogram and Frequency Polygon, Suicide Rates, Large U.S. Cities, 1950*

It should be evident that the procedure in constructing the frequency polygon will parallel that of the histogram in all but the final stages. First, two axes are drawn on the graph paper. Class intervals are then marked ractical purposes only one would ever be constructed. In constructing 'ess than" *CFP* (Figure 4.2.1a), the cumulated frequencies are plotted

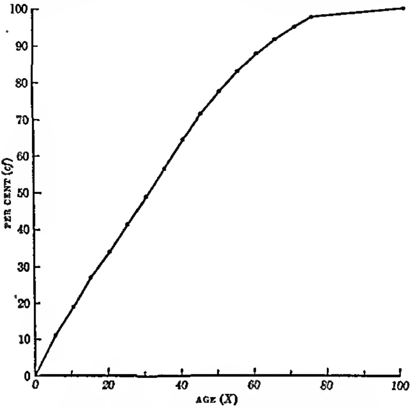


Figure 4.2.1a *"Less than" Cumulative Polygon, Age Distribution, U.S., 1950*

over true upper houndaries of class intervals; whereas in constructing an "or more" graph, the frequencies are plotted over the true lower boundaries (Figure 4.2.1h). This procedure differs from that of the simple

families are less than size 3, and 45 per cent are more than 3, then families of size 3 remain unaccounted for.

When the data are continuous, however, this difficulty does not arise, since the cutting point does not occupy any part of the continuum. Thus, the statement that in 1950, 36 per cent of the United States population was less than 21 years old, and 64 per cent were more than 21, would be precise only if 21.0 is conceived of as the contact point — a necessary statistical convention — between the intervals 20 and 21, which are each one year in width. The problem arises when 21 is treated not as a point, but as an interval of one year, which would be the spontaneous lay interpretation.

The designation employed in this text is sufficiently versatile to satisfy the requirements of the discrete data, and do no violence to the continuous data.

The Nature and Construction of the Frequency Polygon. When items are compressed into a relatively small number of broad intervals, the resulting class frequencies tend to jump in an abrupt manner. It is reasonable to suppose, however, that the progression of class frequencies would be considerably smoother if we employed many relatively small class intervals. It is the function of the *frequency polygon* to provide an approximation of the smooth curve that would presumably emerge if class intervals were made as small as possible and the number of observations were unlimited.

The frequency polygon may be derived from the histogram by simply connecting with straight lines the midpoints of adjacent column tops. Such a conversion is illustrated in Figure 4.1.5a, where a frequency polygon is *superimposed* on the histogram of suicide rates. In practice, only one graph or the other would be presented, depending on whether the emphasis is to be placed on the tabulated class frequencies or on the hypothetical point frequencies, which the frequency polygon provides.

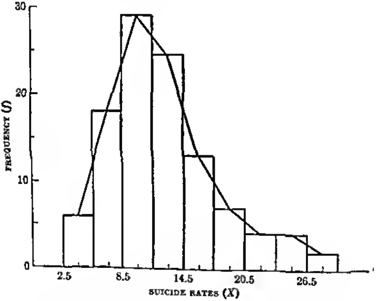


FIGURE 4.1.5a *Histogram and Frequency Polygon, Suicide Rates, Large U.S. Cities, 1950*

It should be evident that the procedure in constructing the frequency polygon will parallel that of the histogram in all but the final stages. First, two axes are drawn on the graph paper. Class intervals are then marked

(b) Combine the last four intervals into two intervals of equal width and show by a heavy line how the appearance of the histogram would be altered.

11. (a) Plot the frequency tables of American and National League 1957 batting averages (Problem 3.1.16) as histograms.
(b) For purposes of comparison, should the histograms be superimposed or plotted separately?

12. (a) Prepare a histogram for the distribution of U.S. families according to size, placing markers at midpoints instead of true class boundaries.
(b) What are the true class boundaries of the intervals?
(c) Justify your choice of an upper boundary to close the graph.

Size of Family	*Per Cent*
2	32.5%
3	22.8
4	20.8
5	12.3
6	5.9
7 or more	5.7
	100.0%

13. (a) Draw histogrâms of the frequency distributions of Indianapolis census tract rentals and educational levels (Problem 3.3.7).
(b) Compare the shape of the two distributions and give a possible interpretation.

14. Construct two frequency distributions: suicide rates (a) east of the Mississippl, and (b) west of the Mississippi (Table 3.1.1a). Convert frequencies to percentages. Plot percentage distributions as frequency polygons on one set of axes. Compare and interpret.

Section Two

The Cumulative Frequency Polygon (CFP), or Ogive

The cumulative frequency table (Table 4.2.1) gives the percentage of cases below (or above) each given class boundary, but not the cumulated frequencies around the innnmerable intermediate values within the interval. Nor can it therefore readily provide those values that correspond to all possible cumulated percentages. Yet information on such intermediate values is frequently required. We may require, for example, the percentage of the population under 21 years of age; or the age such that 50 per cent of the population is younger and older. Although this sort of information could be obtained from the cumulative table by arithmetic interpolation between the class limits, it can be obtained with less effort and sufficient inevitably disturbs the proportionality between column areas and class frequencies. It will always underrepresent the proportion of items in the modal interval.

We should therefore remind ourselves again that no statistical device ever perfectly represents the data on which it plays. The frequency polygon pretends to do no more than provide quickly and economically a rough first approximation of tbe theoretical frequency eurve of values grouped into the smallest possible intervals. For the purpose of comparison, frequency polygons representing different distributions of the same variable may be mutually superimposed without blurring their respective outlines, as would be true of histograms.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Histogram
   Frequency Polygon
   Horisontal Axis
   Vertical Axis
   Frequency Scale
   Variable Scala
   Zero Origin
   Arithmetic Grid Pape1

2. How is the appearance of the histogram changed when the horizontal acale is enlarged in relation to the vertical? The vertical in relation to tha horizontal?

3. What is the advantage of plotting either relative frequencies or absolute frequencies? Is the appearance of the histogram altered when plotting relative frequencies instead of absolute?

4. When is the frequency polygon appropriate for discrete data? (*Hint:* Consider its appropriateness for the smallest possible interval of one and for intervals of increasing width.)

5. State the principles governing the closing of the polygon.

6. If you wished to compare two distributions by superimposing the graph of one on the other, which type would you select: histogram or frequency polygon? Explain.

7. Is the rule that the frequency scale originate with zero as essential in the construction of the polygon as it is for the histogram? Explain.

8. Which figure (histogram or polygon) adheres more closely to the frequency table?

9. Can the polygon be converted as readily into the histogram as the reverse?

10. (a) Represent the frequency distribution of delinquency rates (Table 3.1.7) by a histogram.

all practical purposes only one would ever be constructed. Io constructing a "less than" *CFP* (Figure 4.2.1a), the cumulated frequencies are plotted

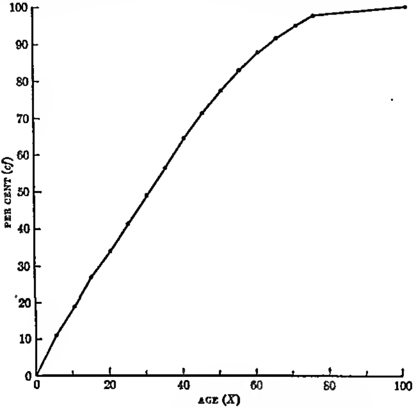


Figure 4.2.1a *"Less than" Cumulative Polygon, Age Distribution, U.S., 1950*

over true upper houndaries of elass intervals; whereas in constructing an "or more" graph, the frequencies are plotted over the true lower boundaries (Figure 4.2.1h). This procedure differs from that of the simple

families are less than size 3, and 45 per cent are more than 3, then families of size 3 remain unaccounted for.

When the data are continuous, however, this difficulty does not arise, since the cutting point does not occupy any part of the continuum. Thus, the statement that in 1950, 36 per cent of the United States population was less than 21 years old, and 64 per cent were more than 21, would be precise only if 21.0 is conceived of as the contact point — a necessary statistical convention — between the intervals 20 and 21, which are each one year in width. The problem arises when 21 is treated not as a point, but as an interval of one year, which would be the spontaneous lay interpretation.

The designation employed in this text is sufficiently versatile to satisfy the requirements of the discrete data, and do no violence to the continuous data.

Table 4.2.1 *Cumulative Age Distribution, U.S., 1950*

Age *	"Less Than" Cumulative Per Cent	"Or More" Cumulative Per Cent
0– 4	10.7 (less than 5)	100.0 (0 or more)
5– 9	19.5 (less than 10)	89.3 (5 or more)
10–14	26.9	80.5
15–19	33.9	73.1
20–24	41.5	66.1
25–29	49.6	58.5
30–34	57.2	50.4
35–39	64.7	42.8
40–44	71.5	35.3
45–49	77.5	28.5
50–54	83.0	22.5
55–59	87.8	17.0
60–64	91.8	12.2
65–69	95.1	8.2
70–74	97.4	4.9
75 and over	100.0 (less than 100)	2.6 (75 or more)

* Age rounded to the last whole year.

Source U S Bureau of the Census, *U.S. Census of the Population: 1950*, Vol. II, *Characteristics of the Population*, Part I, *United States Summary*, U S. Government Printing Office, Washington, D C., 1953

accuracy from the *cumulative frequency polygon* (*CFP*), or *ogive*, as it is sometimes called. "Ogive" is an architectural term for the diagonally curved rib of the Gothic vault, which the cumulative frequency curve often resembles; hence, it was picturesquely so named in 1875 by the English statistician, Francis Galton.

Construction of the CFP The construction begins with plotting axes as for the simple frequency polygon. class intervals are laid off on the base line, and frequencies along the vertical axis. Since frequencies are now cumulative rather than simple, the range of the vertical scale will be equal to the total frequency. For quicker comprehension and ready comparability, the cumulative frequencies are usually expressed as percentages, so that the frequency scale extends from 0 to 100.

Corresponding to the two types of cumulative frequency tabulations — the "less than" and the "or more" * — there are two types of cumulative polygons. But the two graphs provide identical information, so that for

* Some writers employ the terminology "less than" and "more than" — for example, "less than 3" and "more than 3." However, these terms create difficulties when applied to discrete data unless they are treated as continuous. If 33 per cent of all

order to close the tabulation of family incomes (see Table 4.2.2), which reads "15,000 and over," we would have to add literally hundreds of class intervals since the maximum income is in excess of a million dollars. The corresponding *CFP* would be a useless monstrosity.

Applying CFP. The *CFP* permits the frequency distribution to be partitioned at any point whatsoever, according to one's interests and needs. For example, the point that separates the lower 25 per cent and the upper 75 per cent of the items, the so-called *first quartile* (Q_1), may be readily obtained in the following manner (Figure 4.2.1c): from the 25 per cent marker on the frequency axis, run a line parallel to the base line until it intersects the "less than" ogive; from this intersection drop a perpendicu-

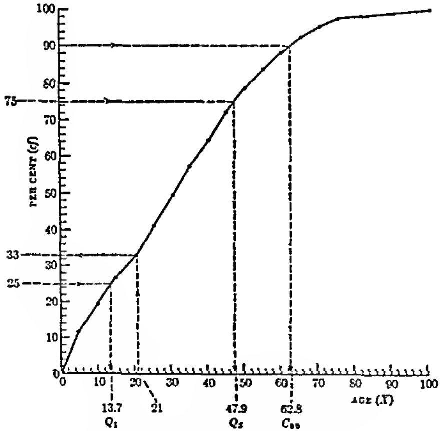


Figure 4.2.1c *"Less Than" Cumulative Polygon, Age Distribution, Selected Centiles and Quartiles*

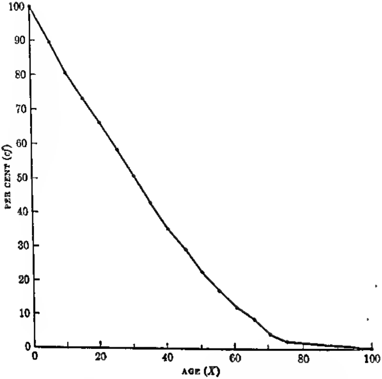


Figure 4.2.1b *"Or More" Cumulative Polygon, Age Distribution, U.S., 1950*

frequency polygon where class midpoints are plotted. Owing to the manner in which scales are laid off, a "less than" *CFP* will begin in the lower left-hand corner of the graph and move diagonally across to the upper right-hand corner; while the "or more" *CFP* will begin in the upper left-hand corner and move diagonally across to the lower right-hand corner.

When one or both ends of the frequency table are open, the *CFP* will not extend over the entire range of the frequency scale and thus will appear incomplete. In that event, it may be possible to close the tabular distribution at some convenient point without doing violence to the data, thereby allowing the ogive to be completed. For example, we may close the United States age distribution at 100 years without injustice to the facts, since only a tiny fraction of all individuals exceed the century mark. To close the cumulative polygon at 100, we would therefore require only several additional five-year intervals on the base line. On the other hand, in

2. (a) What problems arise from unequal intervals in constructing a *CFP?*
   (b) Are these problems identical with those of the simple frequency polygon?

3. (a) Would it be possible to plot a cumulative column diagram, analogous to the histogram? Illustrate.
   (b) When would such a figure be particularly appropriate?
   (c) What are its limitations?

4. Is it possible to deduce the shape of the simple frequency polygon from the cumulative polygon? Explain and illustrate.

5. Explain why the "*more-than, less-than*" *terminology creates difficulties* when applied to discrete data.

Table 4.2.3 *Divorces and Annulments by Number of Years Married, Percentage Distribution, 23 States, 1953*

YEARS MARRIED	NUMBER	PER CENT
Under 1 year	9,260	7.0%
1 year	13,485	10.2
2 years	13,067	9.8
3 years	10,790	8.2
4 years	9,549	7.2
5 years	9,291	7.0
6 years	8,976	6.7
7 years	8,107	6.0
8 years	5,293	4.0
9 years	4,029	3.0
10 years	3,905	2.9
11 years	4,037	3.0
12 years	3,557	2.7
13 years	3,032	2.3
14 years	2,505	1.9
15 years	2,187	1.6
16 years	2,283	1.7
17 years	2,123	1.6
18 years	1,831	1.4
19 years	1,714	1.3
20–24 years	6,137	4.6
25–29 years	4,129	3.1
30–34 years	2,073	1.6
35–39 years	1,007	.7
40 years and over	686	.5
TOTAL	133,103	100.0%

Source: U.S. National Office of Vital Statistics, *Vital Statistics of the U.S., Special Reports, Divorces and Annulments*, Vol. 42, No. 2, 1955, Table 5, U.S. Government Printing Office, Washington, D.C., 1953.

lar to the baseline, which fixes the **first quartile, 13.7.** This is the age below which 25 per cent of the cases lie. The second and third quartiles are found analogously. Centile values may be similarly obtained.* Thus, to find the *90th centile value, C_{90},* we drop a perpendicular to the base line from a point on the curve that represents a "less than" cumulated frequency of 90 per cent; the required value is then read off the base line at the junction point.

By reversing the foregoing procedure, the percentage above (or below) any particular value may be quickly established. We may seek, for instance, the percentage of the population under 21 years of age. We first locate 21 on the base line, at which point we erect a perpendicular to the curve. From that intersection we extend a perpendicular to the frequency axis, where we find that the percentage of persons under 21 years of age is 33 per cent.

It is the ease with which such graphic solutions may be obtained that gives to the cumulative frequency polygon something of an advantage over the cumulative frequency table which would necessitate tedious interpolation.

Table 4.2.2 *Distribution of Families by Income, U.S., 1955*

Family Income	Number of Families (in Thousands)	Per Cent
Under $ 1,000	3,300	7.7%
$ 1,000– 1,999	4,200	9.8
2,000– 2,999	4,700	11.0
3,000– 3,999	6,300	14.6
4,000– 4,999	6,600	15.6
5,000– 5,999	5,400	12.7
6,000– 6,999	4,100	9.5
7,000– 9,999	5,500	12.9
10,000– 14,999	2,100	4.8
15,000 and over	600	1.4
Total	42,800	100.0%

Source U.S. Bureau of the Census, *Current Population Reports*, Series P–60, *Consumer Income*, No 24, U S *Government Printing Office*, Washington, *D.C.*, April, 1957, Table A and Table 2

Questions and Problems

1. Define the following concepts:

Cumulative Frequency Polygon
Ogive
Quartile
Centile
Quantile

* Centiles will be discussed further in Chapter 5, Section 2.

Since the tabulated attributes have only one statistical dimension — namely, the frequency dimension — the corresponding bar chart requires only one scale, the frequency scale. This scale is usually marked off along the horizontal base line. Its range is of course determined by the largest frequency. The attributes are set down by name alongside the bars, and they are easily read when the bars are horizontally placed.

Since there is no continuous quantitative scale (as there is in the histogram) to which the base of the bars must be fitted, the bars may be of any convenient width, placed in any plausible order, and left physically unconnected. But by arranging bars from longest to shortest, the attributes are effectively ranked according to the frequency of their occurrence. The bars are drawn to uniform width merely to maximize the eloquence of the diagram; the dead space between them — ordinarily one-half bar width — serves to give emphasis to the discrete character of qualitative data.

The Pie Chart. As the term suggests, the pie chart (Figure 4.3.1b) is a circular figure in which the areas of the respective slices are drawn proportional to the attribute frequencies. In constructing this chart, percentage frequencies are successively measured with a protractor along the circumference of 360°, or 100 per cent. From these cutting points, we draw radii to the center of the circle, which partition the total area into proportional

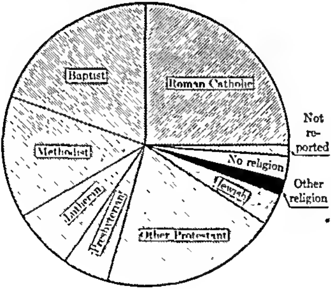


FIGURE 4.3.1b *Pie Chart, Religion Reported, U.S. Population, 1957*

6. (a) Construct "less than" *CFP*'s of the East and West suicide rates on one graph. (Problem 14, p. 85).
   (b) Interpret the difference between the two contours. Are comparisons more effective by means of cumulative or simple frequency graphs?

7. Plot Table 4.2.2 as an "or more" *CFP*.
   (a) How could the lower end be closed?
   (b) How many class intervals would be required to close the upper end?
   (c) From the graph, find the values that enclose the middle 50 per cent of the families.
   (d) What per cent of all families receive $8,000.00 or more?

8. (a) Cumulate the per cent frequencies shown in Table 4.2.3 in both directions and prepare corresponding cumulative polygons on the same set of axes. Verify that the graphs intersect at the 50 per cent marker.
   (b) What percentage of divorces occurs after 10 years of marriage?

SECTION THREE

Graphs of Qualitative Data

The graphic versions of qualitative data differ from those of quantitative data in certain details, but not too drastically in fundamental principles. All graphic portrayals of qualitative data use length, area, or intensity of shading to represent frequency. Only three of the simplest types commonly encountered are here presented: (1) the *bar chart*, (2) the *pie chart*, and (3) the *statistical map*.

The Bar Chart. The bar chart (Figure 4.3.1a) is simply a succession of evenly spaced bars whose lengths are proportional to attribute frequencies.

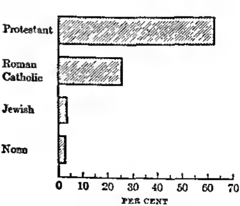


FIGURE 4.3.1a *Bar Chart, Religion Reported, Sample of U.S. Population, 1957*

may be produced by variegated cross-hatebing or by graded concentrations of simple dots. There are innumerable picturesque devices designed to exhibit the diversity among the mapping units. These are described in detail in specialized handbooks on the subject of graphic presentation and are not elaborated here.

The principal problem in geographie mapping is a clean definition of the spatial unit. Unita such as cities or states usually offer few difficulties, since the boundaries of these unita are fixed by law. But sucb units as city blocks, cultural areas, national regions, types of residential districts, and natural ecological areas interpose many ambiguities which must be resolved before their meaning will be adequately conveyed. In spite of sucb problems, the social *base map* is an important tool of social investigation; in fact, it is all but essential to the ecological school of sociology, which focuses on the spatial aspects of social behavior. The social base map consista of those outlines considered essential to the presentation and understanding of tbe plotted social data. Thus, the uneven territorial distribution of alcobolic psychosis is effectively shown in Figure 4.3.2, and poses significant sociological issues in regard to its etiology.

Questions and Problems

1. Define the following concepts:

Bar Chart
Pie Chart
Statistical Map
Social Base Map

2. Describe the distinctive features of graphs of qualitative data.

3. Criticize Figure 4.3.3 as a graph of qualitative data.

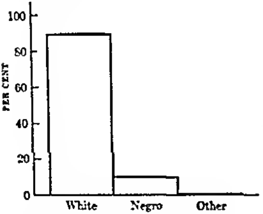


Figure 4.3.3 *"Bar Chart," Racial Composition, U.S., 1950* *

* From U.S. Bureau of the Census, *U.S. Census of the Population: 1950*, Vol. IV, *Special Reports*, Part III, *Nativity and Parentage*, Chapter A, U.S. Government Printing Office, Washington, D.C., 1954.

wedges. The effectiveness of the pie chart, or for that matter all qualitative charts, may often be enhanced by judicious and tasteful shading and coloring.

The Statistical Map. The geographical variation in such items as population composition, marriage and divorce rates, crime rates, and economic indexes is always instructive to the social analyst. The systematic display of such variation may serve the purposes of social policy, as well as of scientific explanation. Thus, the discovery of a relatively high death rate in a certain geographic area, or the concentration of crime and delinquency in specified urban districts, constitutes a first step in uncovering the causes of the phenomenon.

The map is the most natural vehicle for the purpose of setting forth geographic variation. The general principle of statistical mapping is to symbolize varying frequencies by appropriate density of shading, which

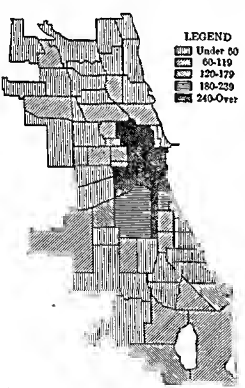


FIGURE 4.32 *Alcoholic Psychosis Rates per 100,000 Adult Population, Chicago, 1922-1931* *

* From R. E. Faris and W. Dunham, *Mental Disorders in Urban Areas*, The University of Chicago Press, Chicago, 1939.

Table 4.3.4 *Per Cent Non-White Population by State, U.S., 1950*

State	Per Cent
New England	
Maine	0.3
New Hampshire	0.2
Vermont	0.1
Massachusetts	1.7
Rhode Island	1.9
Connecticut	2.7
Mid-Atlantic	
New York	6.5
New Jersey	6.7
Pennsylvania	6.1
East North Central	
Ohio	6.5
Indiana	4.5
Illinois	7.6
Michigan	7.1
Wisconsin	1.2
West North Central	
Minnesota	1.0
Iowa	0.8
Missouri	7.6
North Dakota	1.8
South Dakota	3.7
Nebraska	1.8
Kansas	4.0
East South Central	
Kentucky	6.9
Tennessee	16.1
Alabama	32.1
Mississippi	45.4
South Atlantic	
Delaware	13.9
Maryland	16 6
Virginia	22.2
West Virginia	5.8
North Carolina	26.6
South Carolina	38.9
Georgia	30.9
Florida	21.8
West South Central	
Arkansas	22.4
Louisiana	33.0
Oklahoma	9.0
Texas	12.8
Mountain	
Montana	3.2
Idaho	1.2
Wyoming	2.2
Colorado	2.1
New Mexico	7.5
Arizona	12.7
Utah	1.8
Nevada	6.4
Pacific	
Washington	2.7
Oregon	1.6
California	6.3

Source. U.S. Bureau of the Census, *U.S. Census of the Population: 1950*, Vol. II, *Characteristics of the Population*, Table 59, U.S. Government Printing Office, Washington, D C., 1953.

1. Represent Table 4.3.1 by a bar chart.
   (a) Should the width of the bars be adjusted to the unequal age intervals?
   (b) How should the tabulation be closed?
   (c) Explain the declining percentage of high school graduates with increasing age.
   (d) How should bars be ordered?

Table 4.3.1

Percentage of Specified Age Groups Completing High School, U.S., 1950

Age	Per Cent
14–17	2.4
18–24	50.3
25–34	49.1
35–44	37.2
45–54	27.7
55 and over	19.2

Source: U.S. Bureau of the Census, *U.S. Census of the Population: 1950*, Vol. IV, *Special Reports, Education*, U.S. Government Printing Office, Washington, D.C., 1953

Table 4.3.2

Percentage of Specified Income Groups Completing High School, Males 25 and Over, U.S., 1950

Income in 1949	Per Cent
Less than $500	15.0
$ 500–1,999	18.6
2,000–2,999	28 0
3,000–3,999	39.2
4,000–4,999	48.4
5,000–6,999	58.5
7,000 and over	60.5
Not reported	24.5

Source. U.S. Bureau of the Census, *U.S Census of the Population: 1950*, Vol. IV, *Special Reports, Education*, U.S. Government Printing Office, Washington, D.C., 1953.

Table 4.3.3

Rate of Plural Births by Age of Mother, White Population, U.S., 1944

Age of Mother	Rate Per 100,000 Births
10–14	331.1
15–19	549.6
20–24	774.8
25–29	1,019.8
30–34	1,310.7
35–39	1,589.3
40–44	1,292.7
45–49	721.0
50 & over	0.0
All Ages	1,015.2

Source. U.S. National Office of Vital Statistics, *Vital Statistics of the U.S., Special Reports, National Summaries, Plural Birth Statistics, U.S and Each State, 1944*, Vol 25, No. 18, Table D, U.S Government Printing Office, Washington, D C., 1947.

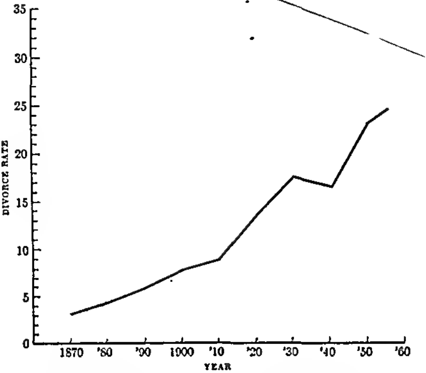


Figure 4.4.1 *Time Chart, Divorces per 100 Marriages, U.S., Selected Years, 1870–1955* *

rule may actually operate to conceal the very fluctuations which the graph hopes to portray. Thus, the trend line on the scale provided in Figure 4.4.2a displays only gentle fluctuations in the annual United States death rate. But such fluctuations are probably more significant than the coarse scale seems to indicate. This discrepancy may be righted by the insertion of a finer scale which throws the annual vacillations into higher relief (Figure 4.4.2b). Now, in order to "preserve" the zero origin and still not enlarge the graph to unwieldy proportions, we introduce a *break* on the vertical axis. A time chart delineating the monthly variation in the United States birth rate uses this little device to good advantage (Figure 4.4.3). What we have done, in effect, is to place the business part of the graph under a microscope in order to magnify the trend line.

However, this is not always an ideal solution. While it is true that by means of this technique relatively small variations may be made more

* From U.S. Bureau of the Census, *Statistical Abstract of the U.S., 1957*, U.S. Government Printing Office, Washington, D.C., 1957, p. 73.

5. (a) Plot Table 4.3.2 as a bar chart.
   (b) Income usually increases with age. As of 1950, the higher income groups show higher educational levels; the older age groups show lower educational levels. How, then, do you account for the increasing percentage of high school graduates in the higher income groups?

6. Prepare a bar chart in vertical position from the data of Table 4.3.3. Draw a line parallel to the horizontal axis through the over-all average of 1,015.2.

7. Prepare a statistical map of the United States to show the percentage of non-white population by state (Table 4.3.4). Use the following shading: states under 15 per cent non-white, *white;* 15–29 per cent, *light gray;* 30–39 per cent, *dark gray;* 40 per cent and over, *black.*

Section Four

Time Graphs

Time series are ordinarily not presented as tables, but are rather converted into graphs — for the simple reason that extended time series are difficult to read as tables, and extremely short series are not significant anyway. The direction, rate, and fluctuation of a trend may be much more readily discerned from a graph than from a series of numbers.

Time graphs are of two prevalent forms: (1) arithmetic charts; and (2) semi-logarithmic (or simply "semi-log") charts.

The Arithmetic Time Chart. This is analogous to the simple frequency graph, except that we plot quantitative observations by intervals of time instead of plotting frequencies on class intervals. The intervals of time are plotted along the base line, and the variable whose fluctuations in time are being recorded is plotted on the vertical axis. In Figure 4.4.1, which exhibits the recent trend in the divorce rate in America, the base line represents the period from 1870 to 1960, and the vertical scale ranges from 0–35 divorces per 100 marriages.

The time scale should be laid out so as to accommodate comfortably *uniform time* intervals in the series; the vertical scale, beginning with the zero origin, should similarly accommodate the largest quantity in the series Entries should, of course, be plotted over the midpoints of the time intervals within which the observation is made.

When the time series is composed of magnitudes or frequencies which move within a narrow range, rigid adherence to the forenamed zero-origin

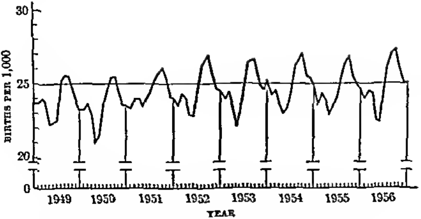


FIGURE 4.4.3 *Time Chart, Monthly Variation in Birth Rate, U.S., 1949–1956* *

of the time series. An illustration of that technique is given in Figure 4.4.4, where successive columns symbolize the number of marriages in the United States each month.

The time graph, which is merely a calendrical series of observations, should not be confused with a bivariate distribution — such as years married and family size — in which the base line represents duration of time. When time is a genuine variable in such a bivariate distribution, it has a true zero point, as in age or any other measure of duration which may be replicated indefinitely. In the time series, however, the markers are no more than arbitrary tags for unique moments in time. Nevertheless, such instants in time are often treated as durations in order to fit a smooth trend line to the observed data.

Multiple Graph. Under some circumstances, it will be possible to plot two or more time series to the same arithmetic scale in order to exhibit more clearly the relation between them. Thus, in Figure 4.4.5, the relation between the divorce rate and the employment rate of married women is rendered more vivid by this device. This graph is readily comprehensible because the two variables have approximately the same range and general location on the scale.

However, this procedure must be undertaken with some caution, *for* it may mislead the reader when variables are not identically located. *In* Figure 4.4.6, the chart may create the impression that in 1945–46 the

* From U.S. National Office of Vital Statistics, *Vital Statistics Special Reports*, Vol. 48, No. 14, 1958, U.S. Government Printing Office, Washington, D.C., 1958.

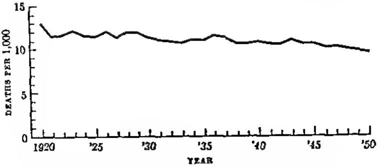


Figure 4.4.2a *Time Chart (Without Break), Death Rate, U.S., 1920–1950* *

visible, it is also true that the movement of the trend line is no longer properly related to the plotted zero origin, which is now an "origin" in name only. Tinkering with the vertical scale will distort the ratios among successive entries and thereby exaggerate or minimize the relative variation However, the practiced eye may very well be able to compensate for such visual distortions.

Instead of a trend line, the time chart may consist of a set of sepaeate columns or bars, whose heights are proportional to the respective values

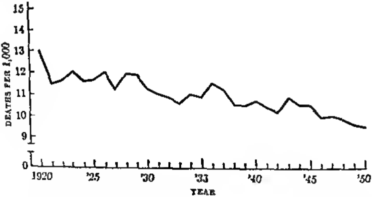


Figure 4.4.2b *Time Chart (With Break), Death Rate, U.S., 1920–1950*

*From U S National Office of Vital Statistics, *Vital Statistics of the U.S., Special Reports, Summary of Mortality Statistics, U.S., 1950,* Vol 37, *No. 8,* Table 1, U S. Government Printing Office, Washington, D C., 1953

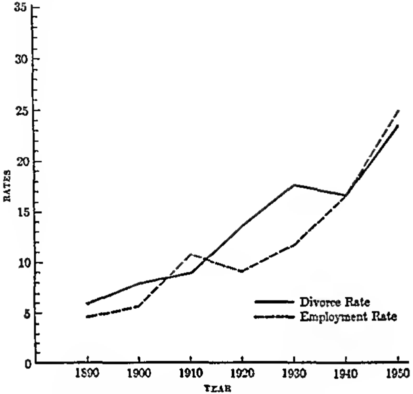


Figure 4.4.5 *Time Chart, Divorce Rate and Percentage of Married Women Gainfully Employed, U.S., Selected Years, 1890–1950* *

Before attempting to graph a time series on semi-log paper, the student should first familiarize himself with the principles underlying the linear arrangement of the guidelines and the markers attached thereto. Initial inspection of the semi-log paper is likely to produce a confused impression, which will be dispelled, however, upon a careful analysis of the logic of the grid scheme.

A typical grid sheet consists of two or more *cycles* which are recognizable as identical linear groupings. The horizontal guidelines within the cycle are not equally spaced, but are so laid down that any specific numerical ratio between two markers always corresponds to the same scale distance: the distance from 2 to 4 is exactly equal to the distance from 1 to 2, 3 to 6, 4½ to 9, 300 to 600, and so on. Consequently, any series of absolute values,

* From U.S. Bureau of the Census, *Statistical Abstract of the U.S., 1957*, U.S. Government Printing Office, Washington, D.C., 1957, p. 73 (adapted).

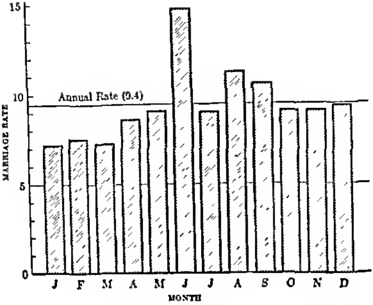


FIGURE 4.A 4 *Time Chart, Monthly Variation in Marriage Rate, U.S., 1956*

marriage rate was rising about 5 times as rapidly as was the divorce rate, whereas, relatively speaking, it was rising only twice as rapidly. This false impression results from the fact that the two sets of data are located at very unequal distances from the zero origin. To compare the rates of change, we have to allow for this locational difference, but this cannot readily be done by the eye. It is the function of the semi-logarithmic chart to transform the scale so as to facilitate comparisons between rates of change

The Semi-logarithmic Chart. In the arithmetic time chart, only the absolute, not the relative, increments of change are directly measurable. Hence, another graphic device is required to measure relative growth and decline; or, to put it more technically, to measure the rate of change in a given variable. This is the *semi-logarithmic*, or *ratio chart*. It reveals at a glance whether the rate of change is constant, or whether that rate itself is changing When the rate of change is constant, as in a geometric series, the graph is a straight line. It is therefore especially appropriate to data of this form, such as the classic idealized Malthusian population series.

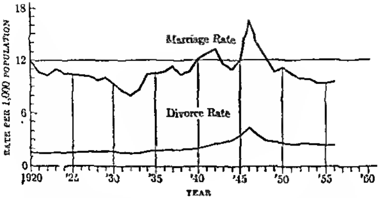


FIGURE 4.4.6 Time Chart, Fluctuation in Marriage and Divorce Rate, U.S., 1920–1956 *

which are plotted on this ratio scale, will automatically show the relative size of successive increments. The student should experimentally familiarize himself with the logarithmic scale by marking off on a strip of paper the distance from, say, 1 to 2 and checking it against other identical ratios, such as 2 to 4, 3 to 6, etc.

Another difference between the ratio and arithmetic scales lies in the numerical origin, which, in the former, must be a value greater than zero. However, for purely practical reasons, it is usually designated as a power of 10: 1, 10, 100; .01, .001, and so on. Since the semi-logarithmic chart measures relative changes, it is impossible to use the zero origin; for any value multiplied by zero yields zero, and hence would never get off the baseline. The Malthusian population could *never multiply if it originated* with zero.

The mathematical foundation of the ratio chart (Figure 4.4.7) is the *principle of common logarithms to the base 10.* However, it will not be necessary for the student at *this time to be intimately familiar* with the procedure of ruling off a logarithmic scale. Prefabricated sheets are commercially available, and the student need merely plot the absolute data from the tabulation.

The markers are *entered on the scale* in multiples of 10 so as to accommodate the values in the time series. Thus, if the range of the data is 30–80, the markers would be written so that 3 = 30, 4 = 40, and so forth. If the range extends from 30 to 120, two cycles would be needed, the second beginning with 100 and *extending* upward to 1,000.

* From U.S. National Office of Vital Statistics, *Vital Statistics of the U.S., Special Reports, National Summaries, Marriages and Divorces*, 1956, Vol. 48, No. 3, U.S. Government Printing Office, Washington, D.C., 1958, Table 1.

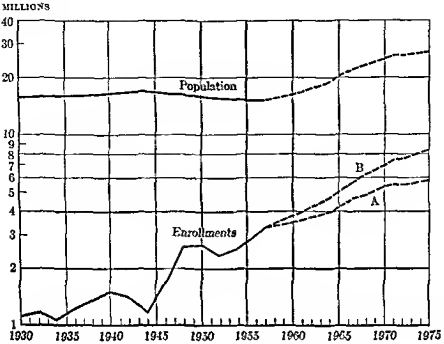


FIGURE 4.4.7 *Semi-logarithmic Chart, College Enrollment and Population, Ages 18–24, U.S., 1930–1957, Projected to 1975* *

In Figure 4.4.7, the college-age population and the actual college enrollments are plotted on the ratio scale. It is easy to read off the semi-log chart that the college-age population has remained almost stationary, but that college enrollments have more than doubled between 1930 and 1950. In this case, the disproportionate increase in enrollment, as related to population growth, stands forth in bold relief. No arithmetic chart, much less a tabulation, could so clearly project the historical facts. Although we could laboriously obtain the absolute figures, we are here not primarily interested in such arithmetic data, nor is the chart adapted to such readings. It is, however, expressly designed to reveal relative changes in the trend. Hence, it is uniquely suited to compare relative shifts either within a given trend, or in two or more trends, irrespective of any divergence in absolute values.

In Figure 4.4.8, the trend of maternal mortality for non-whites is readily compared with the trend for whites, and indicates that the maternal mortality has declined at about the same rate for both color groups. And

* From Metropolitan Life Insurance Co., *Statistical Bulletin*, Vol. 38, August, 1957.

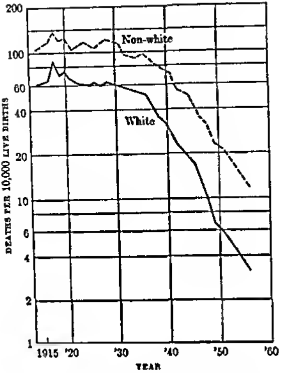


FIGURE 4.4.8 *Semi-logarithmic Chart, Maternal Mortality Rate, White and Non-White, Birth Registration States, 1915–1959* *

even when the trends are located in widely separated belts, it may still be found convenient to plot them in identical cycles, setting up the appropriate multiple scales on both left and right vertical axes — always assuming that the disposition of the curves does not lead to confusion rather than clarity (Exercise 9, p. 108).

As in other comparisons, the juxtaposition of two or more sets of data may suggest "causative" explanations which may be checked by further investigation. However, the sociographic picture, as here portrayed, is only the first stage in the analysis; it supplies an indication of the need for further exploration. Obviously, in the above instance population change does not account for the increase in college enrollment.

* From U.S. National Office of Vital Statistics, *Vital Statistics of the U.S., Special Reports, National Summaries, 1956, Maternal Mortality*, Vol. 48, No. 15, Table A, U.S. Government Printing Office, Washington, D.C., 1958.

The proper interpretation of the semi-log chart assumes an appreciable amount of statistical training. However, even with a modicum of experience a few simple curve-patterns may be identified: (1) a sloping straight line indicates change at a *constant rate;* (2) an upward concave curve, change at an *increasing rate*, and (3) an upward convex pattern, change at a *decreasing rate* (Figure 4.4.9).

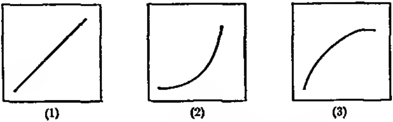


Figure 4.4.9 *Semi-logarithmic Curve Patterns*

Questions and Problems

1. Define the following concepts:
   Zero-Origin Rule
   Scale Break
   Multiple Graph
   Arithmetic Scale
   Arithmetic Time Chart
   Ratio Chart
   Logarithmic Scale
   Cycle
   Constant Rate of Change
   Changing Rate of Change

2. Discuss the hazards involved in plotting two or more time series on the same arithmetic chart.

3. In Figure 4.4.1, in what sense is the trend line inferential rather than descriptive?

4. Construct simple time charts for the time series given in Table 4.4.1.

5. Plot the birth rates of India and the United States (Table 4.4.2) on a single arithmetic time chart. Interpret.

6. Interpret the following semi-log trend lines:
   (a) horizontal line
   (b) steeply sloping straight line
   (c) concave sloping downward

Table 4.4.1

Cancer Mortality Rates, White Males and Females, Ages 1–74, Metropolitan Life Insurance Industrial Policy Holders, 1911–1955

YEAR	DEATHS PER 100,000 **	
	Male	*Female*
1911–1915	66.6	91.4
1916–1920	68.4	89.9
1921–1925	75.4	88.0
1926–1930	80.6	89.0
1931–1935	85.8	89.0
1936–1940	86.8	89.0
1941–1945	85.4	82.5
1946–1950 *	92.3	79.9
1951–1955 *	97.5	75.6

* Based on 6th Revision of the International List of Causes of Death.

** Death rates standardized for age.

Sources Metropolitan Life Insurance Company, *Statistical Bulletin*, July 1945; E. A. Lew and M. Spiegelman, "The Mortality Experience of Industrial Policyholders, 1950–1955," *Society of Actuaries Transactions*, Vol. 9, May 1957.

7. Why is the semi-log chart peculiarly adapted to the comparison of rates of growth?

8. Why is the Malthusian population series called a geometric series?

9 Plot the time series shown in Table 4.4.3 on semi-log paper, using appropriate multiple scales.

Table 4.4.2

Average Birth Rate, Five Year Periods, 1916–1945, India and U S

YEARS	BIRTH RATES	
	India	*U.S.*
1916–1920	34.7	27.9
1921–1925	33.0	26.3
1926–1930	33.8	22.5
1931–1935	34.6	19.2
1936–1940	33.5	18.9
1941–1945	28.3	21.4

Sources: U.S. Department of Commerce, *Vital Statistics — Special Reports, Summary of Natality Statistics*, 1951. U S. Government Printing Office, Washington, D C., 1954, Table 1; Kingsley Davis, *The Population of India and Pakistan*, Princeton University Press, Princeton, N.J., 1951, Table 18, p. 68.

Table 4.4.3

Census of Population (in Thousands), California, Kansas, and San Diego, Decennial Years, 1890–1950

Year	California	Kansas	San Diego
1890	1,213	1,428	16
1900	1,485	1,470	18
1910	2,378	1,691	40
1920	3,427	1,769	74
1930	5,677	1,881	148
1940	6,907	1,801	203
1950	10,586	1,905	334

Source: U.S. Bureau of the Census, *Statistical Abstract of U.S., 1956*, U.S. Government Printing Office, Washington, D.C., 1956, *Tables 8 and 11*

Selected References

Croxton, Frederick E., and D. J. Cowden, *Applied General Statistics*. Second edition. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1956. Chapters 4–6.

Modley, Rudolf, and Dyno Lowenstein, *Pictographs and Graphs*. Harper & Brothers, New York, 1952.

Schmid, Calvin F., *Handbook of Graphic Presentation*. The Ronald Press Company, New York, 1954.

Spear, Mary Eleanor, *Charting Statistics*. McGraw-Hill Book Company, Inc., New York, 1952.

Averages 5

Section One

The Mode

Measures of Location. The frequency distributions of quantitative data which have been previously analyzed constitute condensations of large masses of observations. Basically, they are nothing more than series of values deployed on a continuous scale. We may therefore declare that the items ere *located* on a segment of the scale. Such a description becomes particularly meaningful when we compare two or more distributions which are located in different regions of the same scale. Thus, when husbands end wives are classified by age and located on the same age scale, the husbands, who are usually older, occupy the upper end of the scale, and wives, being younger, occupy the lower end. Similarly, whites and nonwhites in the United States ere differently located on the wage continuum (Figure 5.1.1). It is clear that the statistical concept, *location*, is concerned only with quantitative variables, and is not applicable to qualitative data, which have no scale to be located on.

But when making these comparisons, it is not always practical to quote or depict the full distribution, however compactly it may be presented in tabular or graphic form. For many purposes, the complete table is unnecessarily detailed, too cumbersome to manipulate, and not always readily comparable to other tables of analogous data. In accordance with the general function of statistics to simplify large masses of data, we need a more condensed statement that will (1) provide information on the locational value in which we are interested, (2) eliminate those values which are at the moment irrelevant, and (3) will still faithfully represent the totality with reasonable efficiency.

The limit of such efficient condensation would obviously be the reduc-

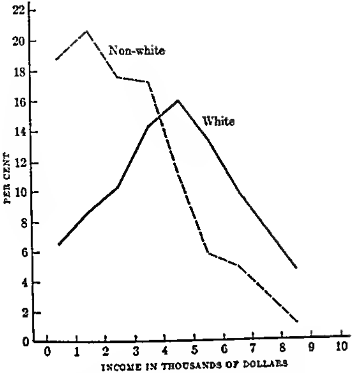


FIGURE 5.1.1 *Annual Family Income, Percentage Distribution, White and Non-White, Urban U.S., 1955* *

tion of the multitude of items to one single value which would, in some way, represent the entire aggregate. But it is clear that no single value is sufficiently versatile to reflect every characteristic of location of a distribution; instead, it can reflect only one feature of it. The representative value is not, therefore, a replica or miniature of the total; it is rather a selected value of limited utility which will "do the work" for the totality. When this task is to fix the location of the distribution along the scale, then the value is called a *measure of location.*

Any value in the distribution could serve to represent the totality, if we knew the position of that item in the full array. Hence, we must analyze all of the items in the aggregate in order to assess the representativeness of any one. But, in practice, not all values will be equally serviceable, even though their representativeness has been determined; rather three types of values have been found most convenient and useful to

* From U.S. Bureau of the Census, *Current Population Reports, Series P-60, No. 24, Consumer Income,* U.S. Government Printing Office, Washington, D.C., 1957.

abstract from a tabulation: (1) the *maximum*, (2) the *minimum*, and (3) the central or typical values, known as *averages*.

The Maximum. There are occasions when the maximum value in a distribution is the *only* useful one. In the fluctuations of traffic volume and weight, it is the peak load which must be safely accommodated on the highway or bridge; the "average" is not a satisfactory guide, since the bridge built for the average load would collapse under the maximum traffic. Similarly, the capacity of schools, hospitals, and other institutions is planned for the approximate maximum, not for the anticipated average patronage. *Middle values are not safe for such purposes.* The maximum, therefore, as a measure of location, constitutes that value below which lie the temporarily irrelevant values.

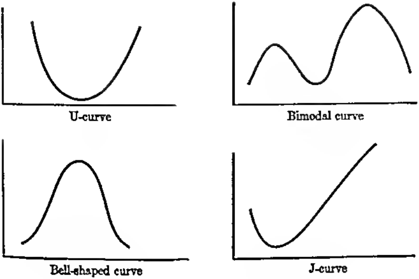
The Minimum. Many problems of social policy are resolved by selecting the minimum value of a distribution as a working norm. Obviously, the minimum is that value above which all other values in the array may be found. From the distribution of incomes of healthy families, the minimum income is selected for legislated welfare norms. From the hypothetical distribution of mature individuals by age, the minimum age of eligibility is set for marriage, army service, voting, and other social responsibilities. A prospective college student may be interested only in the minimum of the array of college expenditures and use it as a guide in making his plans The other values are disregarded.

Neither the maximum nor minimum offers serious computational or statistical complications either in reckoning or interpretation. They are simple in conception; hence, they require no extended discussion.

The Concept of Average. Any value between the extremes could be selected as a locational measure. However, the most common intermediate locational value, and in general the most useful, is an *average*. This is generally a central value around which the distribution seems to cluster. This apparent tendency of many statistical aggregates to concentrate around a center is often termed "central tendency," the value at this center being the *measure of central tendency*, more commonly called *average*.

This functional center, however, is not necessarily identical with the middle region of the range of observed data. The region of concentration may be either near the midpoint of the range or at a considerable distance from it. Thus, the distribution of intelligence scores resembles a bell-shaped curve which is centered on the midpoint of the range. On the other hand, the distribution of wages, sizes of cities, and other social variables may present U-curves or J-curves whose centers are displaced toward the edges of the range, as in the sketches in Figure 5.1.2.

As in common usage, so also in statistical language, the concept "aver-


U-curve

Bimodal curve

Bell-shaped curve

J-curve

Figure 5.1.2 *Types of Frequency Curves*

age" connotes the typical, the ordinary, and expected. Like most other statistical measures, the average has its roots in common experience and is indispensable to daily discourse. Every layman summarizes and simplifies the whole range of his experiences by speaking quite casually about the average voter, the average family, the average student, a batting average, or may refer to almost any experience as "just average." The wide variety of traits which are reducible to an average is shown in such popular descriptions as that of the "average" American male who "stands five feet nine inches, weighs 158, prefers brunettes, baseball, beefsteak, and French fries, and thinks the ability to run a home efficiently is the most important quality of a wife."

There is, of course, no male who is average in every respect: the person of average height will not necessarily be a person of average intelligence or average handsomeness. Therefore, the popular claim that the "average person" does not exist is completely justified. Such unrealistic expressions as "average school" and "average Negro" do not conform to the stricter statistical concept which applies only to a series of measures on a single variable.

Nevertheless, in whatever manner he employs the concept, the layman unconsciously implies what every statistician explicitly recognizes: that the average is a kind of norm around which the values tend to vary. But the difference between the layman and the professional is that the latter requires more precision than is provided by informal folk usage. He therefore devises mathematical procedures to measure the average, and thereby restricts himself to variables whose central tendency is reducible to some form of quantification. Since there are various types of central tendency, be necessarily formulates various averages, each answering to the requirements of a given problem. Of these numerous averages, only three are here discussed: the *mode*, the *median*, and tbe *arithmetic mean*. The arithmetic mean is the best known and most widely used in statistical work; however, tbe mode and the median are conceptually more elementary and for tbat reason will be taken up first.

The Mode, or Probability Average

The mode is simply the most frequently recurring value in an ordered distribution; it is that value where the concentration of items is most dense. Etymologically, it is related to the notion of the prevailing fashion of dress or etiquette to which a majority of a given social class would be expected to conform. Hence, the mode (*Mo*) may also be defined as the most probable value, and therefore distinctively labeled the *probability average*.

An examination of popular expressions suggests that the mode is, in fact, often implied by the concept "average." Such usage stems, in part, from the fact tbat attributes as well as variates may take on predominant frequencies in a *series of observations* When a politician refers to the "average voter" as wanting his interests protected, he usually means that a majority of voters are motivated by self-interest; or when a waitress remarks that the "average customer" does not drink his coffee black, she is likely to mean that most restaurant patrons do not drink black coffee. Restaurant patrons and voters, in their turn, speak of the "average waitress" and the "average politician."

Statistically speaking, the mode is the most likely value to be met with in a series of recorded observations. Suppose, for example, a citizen of Peoria were to estimate the suicide rate of his community, baving in his possession only the table of suicide rates for the group of 107 cities (Table 3.1.1d). His best guess in this instance would be the most frequently occurring rate, somewhere between 8.5 and 11.5. By selecting the interval with the largest single frequency (28 per cent), to that extent, he would more likely be right than wrong. The probability of success would depend, of course, on the degree of preponderance of the most frequently occurring value *in the given distribution*. If, instead of 28 per cent, the preponderant frequency were 15 per cent or 75 per cent of the total, tbe probability of being correct would be diminished or increased accordingly. But the modal value would still remain the same, irrespective of the size of the plurality of the most numerous category. Thus, the mode is the most probable value one would encounter, but it does not reveal the degree of that probability. Clearly, if the frequencies of every value were identical there would be no mode.

Calculation of the Mode. Since the mode is the most frequently occurring value, it is obviously necessary to count the number of occurrences of each value. In the case of continuous data, it is conceivable that empirical measurements would be so minute and discriminating that no two cases in the set would be found to be identical in value. An examination of the 107 suicide rates reveals that very few rates are indeed identical when they are measured to two decimal places. If, however, a little precision is sacrificed by rounding and grouping, a mode very quickly emerges. In fact, a mode is impossible without grouping, since without grouping there would be no frequencies. It goes without saying that class intervals must be identical in width; otherwise one could, by making a particular interval large enough, obtain an apparent mode of almost any desired value — a clearly meaningless result.

There are, then, two distinct steps in the determination of the mode: (1) locate the predominant, or *modal*, frequency, and (2) find the value corresponding to that frequency. This is illustrated in Tables 5.1.1a and 5.1.1b.

Table 5.1.1a

Computation of Crude Mode

Class Interval	*f*
$1–2	5
3–4	6
5–6	2
	13
Mode = Mdpt. of $2.50 – 4.50 = $3.50	

Table 5.1.1b

Computation of Crude Mode (Data Regrouped)

Class Interval	*f*
$1	2
2	3
3	5
4	1
5	1
6	1
	13
Mode = Mdpt. of $2.50 – 3.50 = $3.00	

In Table 5.1.Ia, the highest frequency is 6, and the modal value, or *crude mode*, is \$3.50, which is the midpoint of that interval, 2.5–4.5. However, with the finer grouping of Table 5.1.1b, the highest frequency is 5, and the midpoint of the corresponding interval is \$3.00. It immediately becomes evident that, since there is some flexibility in the choice of class width, there will ensue a certain instability in the resulting mode.

In geneml, statisticians do not like such instability of values, resulting from an adventitious circumstance like the selected size of an interval. Instability tends to discredit the very validity of any measure. To circumvent this dilemma, two alternatives are available: (1) abandon that average and use another, or (2) refine the crude mode by a method of calculation which would reduce its instability. But with all its admitted instability, the mode cannot be abandoned in favor of another average if it is the mode that one wants. Hence, we must turn our attention to the refinement of that measure.

The crudest method, discussed and illustrated above, designates the midpoint of the interval of the most populous class; it ignores the immediately adjacent intervals and their frequencies. But these adjacent intervals would have affected the value of the mode if the class boundaries had been placed otherwise. Hence, a more sensitive formula has been designed to reflect the "pulling power" of these adjacent frequencies, and to allow the operation of their force. Its application presumably produces the result that would be obtained if class intervals were made progressively smaller in order to secure a mora etable and accurate approximation of the point of greatest density.

The Difference Method. This method of rafining the mode proceeds: (I) by calculating the differences between the modal frequency and the respective adjacent class frequencies; (2) by calculating the ratio of one of these differences (usually the next lower) to the sum of the two differences; (3) by applying this proportion to the modal class width; this result, when (4) added to the true lower boundary of the modal interval, serves to fix the value of the refined mode. The formula:

$$Mo = L + \left(\frac{D_1}{D_1 + D_2} \times i\right) \tag{5.1.I}$$

where Mo = refined mode

L = the true lower limit of the modal interval

D_1 = the difference between the modal frequency and the frequency of the next lower interval

D_2 = the difference between the modal frequency and the frequency of the next higher interval

i = the class interval

Applying this formula to Tables 5.1.1a and 5.1.1b we obtain the following results:

(a)

$$Mo = \$2.50 + \left(\frac{1}{1+4} \times 2\right)$$
$$= \$2.50 + .40$$
$$= \$2.90$$

(b)

$$Mo = \$2.50 + \left(\frac{2}{2+4} \times 1\right)$$
$$= \$2.50 + .33$$
$$= \$2.83$$

The mode has now been "pulled" from \$3.50 and \$3.00 to \$2.90 and \$2.83 respectively, and thus displays a marked gain in stability. Where before the modes differed by 50 cents, they now differ by only 7 cents. The influence of grouping is reduced by allowing the adjacent frequencies to participate in the result.

Graphic Method. The graphic version of the difference method is illustrated in Figures 5.1.3a and 5.1.3b. We first convert each tabulation into a histogram. The procedure continues by connecting the true limits of the columns adjacent to the modal column by diagonals in the manner

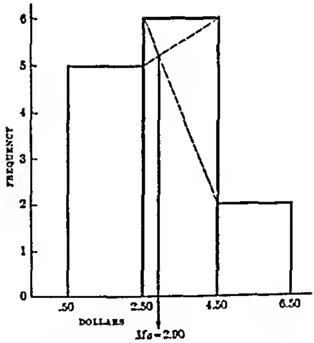


Figure 5.1.3a *Difference Method for Determining Refined Mode (Based on Table 5.1.1a)*

In Table 5.1.1a, the highest frequency is 6, and the modal value, or *crude mode*, is $3.50, which is the midpoint of that interval, 2.5–4.5. However, with the finer grouping of Table 5.1.1b, the highest lrequency is 5, and the midpoint of the corresponding interval is $3.00. It immediately becomes evident that, since there is some flexibility in the choice of class width, there will ensue a certain instability in the resulting mode.

In general, statisticians do not like such instability of values, resulting lrom an adventitious circumstance like tbe selected size ol an interval. Instability tends to discredit the very validity of any measure. To circumvent this dilemma, two alternatives are available: (1) abandon that average and use another, or (2) refine the crude mode by a method ol calculation which would reduce its instability. But with all its admitted instability, the mode cannot be abandoned in lavor of another average if it is the mode that one wants. Hence, we must turn our attention to tbe refinement of that measure.

The crudest metbod, discussed and illustrated above, designates the midpoint of the interval ol tbe most populous class; it ignores the immediately adjacent intervale and their frequencies. But these adjacent intervals would have affected the value of the mode if tbe class boundaries had been placed otherwise. Hence, a more eensitive formula has been designed to reflect the "pulling power" of these adjacent frequencies, end to allow tbe operation ol their force. Its application presumably produces tbe result tbat would be obtained if class intervals were made progressively smaller in order to secure a more stable and accurate approximation of tbe point of greatest density.

The Difference Method. This method of refining the mode proceeds: (1) by calculating tbe differences between the modal frequency and the respective adjacent class frequencies; (2) by calculating the ratio of one of these differences (usually the next lower) to the sum of tbe two differences; (3) by appiying this proportion to the modal class width; tbis result, when (4) added to the true lower boundary of the modal interval, serves to fix the value of the refined mode. The formula:

$$Mo = L + \left(\frac{D_1}{D_1 + D_2} \times i\right) \qquad (5.1.1)$$

where Mo = refined mode
L = the true lower limit of the modal interval
D_1 = the difference between the modal lrequency and the frequency of the next lower interval
D_2 = the difference between the modal frequency and the frequency of the next higher interval
i = the class interval

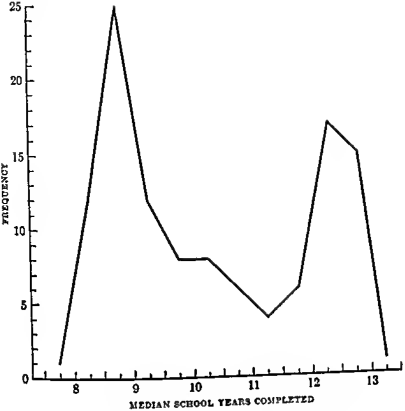


FIGURE 5.1.4. *Bimodal Frequency Distribution, Median School Years Completed, 115 Census Tracts, Indianapolis, 1950*

modes probably reflect prevalent terminal points in school attendance: a large number leave school immediately after attaining the legal minimum age; another large concentration take their leave after they have completed high school. When confronted with such bimodality, the worker is often called upon to disentangle the populations which caused it; or, failing that, he would have to accept the dual modality as a valid description of a single population.

Evaluation of the Mode. While the mode would appear to recommend itself for measuring representativeness — what is more representative than the most frequent value? — some of its characteristics disqualify it for any but the simplest purposes. First, it cannot be subsequently manipulated by algebraic rules because of its own derivation — a point which will become clearer when we discuss the mean; (2) it is influenced by the width

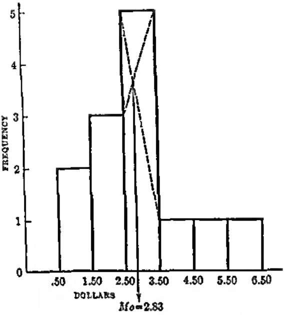


FIGURE 5.1.3b *Difference Method for Determining Refined Mode (Based on Table 5.1.1b)*

shown. Then, from the point where these diagonals intersect, we drop a perpendicular to the base line. This perpendicular cuts the base line at the refined mode, which is simply read off the scale. Such a graphic presentation shows more clearly than does the arithmetic calculation how the discrepancy between the two differences determines the position of the mode: the greater that discrepancy, the greater the displacement from the crude mode (the midpoint of the modal interval). When there is no discrepancy, there is of course no displacement at all.

Bimodality. Some distributions display two concentrations or humps, and are therefore called *bimodal* to distinguish them from *unimodal* distributions This bimodality of a given distribution is generally the result of amalgamating two or more populations which have markedly different locations on the scale. Thus, the bimodality in the frequency polygon of adult heights is due to the consolidation of groups of males and females, who are characterized by two different sets of heights. A graphic example of bimodality is furnished by the frequency distribution of the average schooling for 115 Indianapolis census tracts (Figure 5.1.4). These two

4. In the text, it was stated that it is impossible to determine the degree of modality from the mode. Does this statement still hold true when the mode is corrected according to the difference formula?

5. In computing the refined mode, what assumption is made concerning the distribution of the items in the modal class interval?

6. List two instances of bimodal frequency distributions of social data.

7. For what kind of data is the mode the only possible average?

8. Can the degree of modality be approximately inferred from the value of the refined mode? For example, if the true modal family size is 2.9, does this not indicate that the most frequently appearing family size is 2, but that it does not predominate much above the frequency of size 3?

9. Compute the refined mode for the distribution of suicide rates (Table 3.1.1d) according to Formula 5.1.1.

10. Compute the refined mode of the distribution of delinquency rates (Problem 11, p. 47) by Formula 5.1.1.

Section Two

The Median, or Position Average

The Principle of Ordinal Position. In any given array, every item holds a certain rank, be that the first, second, tenth, or seventy-fifth. Obviously, any particular numerical rank takes its meaning and significance from the total number of ranks there are. A rank of 10 in a series of 100 is relatively higher than a rank of 10 in a group of 20.

The point which cuts the array into two equal divisions, so that exactly one-half $\left(\frac{N}{2}\right)$ of the items are below, and one-half are above that point, is called the *median*. The "average" student, for example, usually considers himself near the middle of the array with about 50 per cent of the students below and above. Although, according to this method of reckoning, the average student would also display the most frequent score (i.e., the mode), the focus here is on the position in the rank order, rather than on his membership in the modal group. Therefore, the student who stands exactly in the middle of the lot, with a grade of 80, for instance, is thought of as the median student, and his grade or score is the median value, or simply the median. In 1950, the median age of the total United States population was 30.4 years, signifying that one half of the population was older, and one half younger than that age. Since the median clearly denotes position in a sequence of values, it is often referred to as a *position average*.

of the class interval employed, and therefore to that extent lacks stability; and (3) it is unable to show its degree of modality and is in that sense nonspecific. We may legitimately raise a question about the serviceability of a measure which tells you the most frequent value, but not its relative weight in the distribution, i e., the *degree of modality*. It gives you the most probable value, but offers no clue to how probable that value is. If this is required, the whole distribution must again be examined. Finally, as to how large the predominant frequency should be in order to dignify it as a modal frequency, there is no rule but good practical sense.

Any criticism that might be leveled against the mode is not to deny that this measure is frequently quite useful. A housing administrator, planning a building program, will be interested in the predominant size of the families to be domiciled. He will not employ the mean size of all families, nor the median, but will estimate the most frequent size which is likely to occur among his renters. A young law graduate, in embarking on his career, will not estimate his probable income on the prevailing mean computed from all known incomes, but rather on the predominant income for his class. Clothing merchants will stock their stores according to their expectations of modal patronage. The fact that the degree of modality may not be readily available does not alter the need for information on the predominant frequencies of the items of interest. The mode is, furthermore, the only average which is applicable to qualitative variables.

Although the technical terminology may not be employed by the housing administrator, the aspiring lawyer, or the clothing merchant, his conceptual focus is still on the mode, which he informally and vaguely labels the "average."

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Average
   Measure of Location
   Probability Average
   Modal Frequency
   Crude Mode
   Refined Mode
   Degree of Modality
   Difference Method
   Bimodality

2. Under what circumstance does the refined mode equal the crude mode?

3. For a given N, what would be the effect on degree of modality of the following circumstances?
   (a) a wider range of data with more concentration in the tails
   (b) a narrower range, with less concentration in the tails

number of items below and above the median would now be three. Counting off three items from either end, we would meet half-way between ranks 3 and 4, or at 3.5, which is the true upper limit of the third and the true lower limit of the fourth rank. The corresponding point in the array of values is a point midway between 5 and 7. Since this interval extends from 4.5 to 7.5, the median point is exactly 6.00. Actually, any value between the midpoints 5 and 7 would satisfy the definition of the median, since the median is a point such that half of the ranked items are above and half below it. It is conventional, however, to locate the median exactly halfway between the values, a placement which is consistent with the even division of ranks.*

Procedure in Grouped Data. Most statistical problems are not as simple as those cited above; in fact, such simple problems are quite unrealistic, for statistical investigations generally involve large masses of grouped data. Although in such cases, the procedure for computing the median is fundamentally the same as that for ungrouped data, it must necessarily accommodate itself to the fact that the median will almost always lie somewhere within a class interval. This median interval itself can be readily identified from the cumulative frequency table, but it is still necessary to establish the median point within that interval.

The cumulative frequency distribution of Table 5.2.2a will demonstrate the method of determining the median age of the population of the United States in 1950. (Observe that the frequencies have been rounded to the nearest million, which is a type of coding that will not significantly influence the median and will save much tedious work.)

Conforming to the above directive, we first divide the total frequency into halves: $\frac{150}{2} = 75$. We then locate the class interval of the $\frac{N}{2}$th, or the 75th case, by counting from either end. Cumulating from the lower end, the 75th case is obviously not among the first 16 cases, nor is it among the first 62 cases, which represents the cumulation of the first three frequencies. The fourth cumulation of 86 overshoots the mark by a sub-

* In some texts, the median of ungrouped data is located according to the formula, $\frac{N+1}{2}$, which may be called the *method of the middle case*. This would yield for Table 5.2 1a the median rank of 3, and for Table 5 2.1b, the rank of 3.5. The corresponding median values would be 5 and 6 respectively. Although the results are identical with those above, the method is not recommended. It has led to the misconception that $\frac{N+1}{2}$ items are below (and above) the median, which is contrary to the fundamental definition. Moreover, this formula is inapplicable to grouped data. There seems to be no compelling reason for separate formulas for ungrouped and grouped data, since these data differ only in tabular appearance. We therefore adhere uniformly to the $\frac{N}{2}$ formula.

Calculation of the Median. Since the median divides the range into two parts, with each division containing exactly 50 per cent of the items, it is nothing more than the common true boundary between those two segments. There are therefore three essential steps in the determination of the median: (1) *arranging the values in rank order*; (2) *finding the position of the* $\frac{N}{2}$th item; (3) determining the value of that item.

With ungrouped data, involving only a few cases, the median could be found by inspecting the array. Let us suppose that five persons have respectively 5, 2, 7, 4, and 10 dollars each. In order to find the median, we arrange the values in order of magnitude, and designate the position, or rank of each, as shown in Table 5.2.1a. According to the basic definition, the number of items above and below the median is $\frac{N}{2}$, in this case 2½ items. Hence, starting at either end, we count through the $\frac{N}{2}$th, or 2½th rank, which places the half-way point exactly in the middle of the third rank — it bisects the third rank.* The corresponding value, or the median, would be the midpoint of 5, or exactly 5.00.

Table 5.2.1a

Computation of Median, Odd Number of Items

X	Rank	
$ 2	1	2½ ranks below
4	2	
5	3	*Median Point*
7	4	2½ ranks above
10	5	

When the array contains an *even* number of items, a short additional step is required. If, for example, an item of $13 were added to the array of Table 5.2.1a, we would have the situation shown in Table 5.2.1b. The

Table 5.2.1b

Computation of Median, Even Number of Items

X	Rank	
$ 2	1*	3 ranks below
4	2	
5	3	
		Median Point
7	4	3 ranks above
10	5	
13	6	

* We should remind the reader that it is quite conventional to treat discrete data, such as ranks, as if they were continuous.

Nor would the result differ if we operated on percentage frequencies, instead of absolute frequencies. Since $\frac{N}{2}$ is necessarily 50 when the calculation is performed on percentages, the formula becomes:

$$Md = L + \left(\frac{50 - cf}{f} \times i\right)$$

it being understood tbat f and cf stand for the appropriate percentage frequencies (see Table 5.2.2b). We note finally that, as in the above example, the frequency table need not be closed in order to obtain the median.

Table 5.2.2b

Computation of Median Age, U.S. Population, 1950

Age	Per Cent	Cumulated Percentages
	10.7	10.7
	16.1	26.8
	14.8	41.6
25–34	15.8	57.4
35–44	14.2	
45–54	11.5	
55–64	8.8	
65 and over	8.1	
	100.0%	

$$Md = 25.0 + \left(\frac{50 - 41.6}{15.8} \times 10\right)$$
$$= 25.0 + \left(\frac{8.4}{15.8} \times 10\right)$$
$$= 30.3 \text{ years}$$

Median of Discrete Data. Some writers would limit the use of the median to continuous data, for the reason that discrete data by definition cannot be fractionated as would be required for the median. But such a prohibition does not seem too feasible. In the discrete tabulation of Indiana families by size (Table 5.2.3), there is no observed family size such that exactly 50 per cent of the families are larger, and 50 per cent smaller. We find that 34 per cent of all families consist of 2 persons or less, and 56 per cent consist of 3 persons or less. Hence, some of the three-person families lie below the 50 per cent point, and the remainder lie above it — a circumstance which seems to exclude the possibility of a clean median family size. In this situation, must we then abandon tbe median, or may we pragmatically treat the data as continuous and proceed to accept a fractional

Table 5.2.2a

Population by Age, Cumulative Distribution, U.S., 1950

AGE	f (IN MILLIONS)	cf
Under 5	16	16
5–14	24	40
15–24	22	62
25–34	24	86
35–44	22	108
45–54	17	125
55–64	13	138
65 and over	12	150
TOTAL	150	

Source: U.S. Bureau of the Census, *U.S. Census of the Population- 1950*, Vol. II, *Characteristics of the Population*, Part 1. U.S. Summary, Table 39, U.S. Government Printing Office, Washington, D.C., 1953.

stantial margin. The 75th case is therefore somewhere within the class frequency of 24, which is thought of as being spread uniformly throughout the interval, 25–34. Having accounted for 62 cases, we need to pass through 13 additional items (75 minus 62) of the 24 in order to include the desired item, and to establish the position of the median. Starting at its true lower limit, we must, therefore, penetrate the fourth class interval $\frac{13}{24}$th, or 54 per cent, of its width. Since the class interval is 10, the distance still to go to reach the median is .54 × 10, which is 5.4 years. Adding this amount to the true lower limit of the class interval (25.0), the median is found to be 30.4 years.

The entire procedure may be compactly written as follows:

$$Md = L + \left(\frac{\frac{N}{2} - cf}{f} \times i\right) \tag{5.2.1}$$

where Md = median

L = the true lower limit of the class interval in which the median is located

$\frac{N}{2}$ = one half of the total frequency

cf = the cumulated frequency up to the median class interval

f = the frequency of the median class interval

i = class width

Substituting in this formula, we arrive at exactly the same result as before:

$$\begin{aligned} Md &= 25.0 + \left(\frac{75 - 62}{24} \times 10\right) \\ &= 25.0 + 5.4 \\ &= 30.4 \end{aligned}$$

even hundredth. We may state that Mr. Jones is in the upper half of the income distribution, but to place him in the highest 1 per cent is certainly more informative, although both statements may be true. For such added discrimination, smaller subdivisions are required. The computation of *quartiles, deciles,* and *centiles,* which divide the array into fourths, tenths, and hundredths, respectively, is carried out according to the same principle as in the case of the median, except that the appropriate frequency, or percentage, of items below the point in question is substituted for $\frac{N}{2}$ in the formula given above. For example, to find the point below which the lowest quarter of the cases fall, we replace $\frac{N}{2}$ by $\frac{N}{4}$. Thus, the first quartile in the age distribution shown in Table 5.2.2a is:

$$\begin{aligned} Q_1 &= \mathrm{L} + \left(\frac{\frac{N}{4} - cf}{f} \times i\right) \\ &= 5.0 + \left(\frac{37.5 - 16}{24} \times 10\right) \\ &= 5.0 + 9.0 \\ &= 14 \text{ years} \end{aligned}$$

If we wish to locate the point below which 75 per cent of the items fall (or Q_3), we make the following substitution in the formula:

$$Q_3 = L + \left(\frac{\frac{3N}{4} - cf}{f} \times i\right)$$

The 90th centile (or C_{90}) would be found by the formula:

$$C_{90} = L + \left(\frac{\frac{90N}{100} - cf}{f} \times i\right)$$

Quantiles as Standardized Measures. The median, quartiles, quintiles, deciles, and centiles, which by their definition indicate the proportion of items that are located below or above a given value, are collectively referred to as *quantiles.* As has been demonstrated in this section, quantiles may be used to fix the relative position of any given value in its array. A weight of 180 pounds may be located at the 90th centile (C_{90} = 180 lbs.) which places that weight in a position such that 10 per cent of the population are above that weight, and 99 per cent below it. Similarly, an age of 62, an IQ of 120, a height of 5 ft., 9 in., an income of $10,000, a suicide rate of 22 per 100,000—all may conceivably be located at exactly the same abstract centile point.

It is now evident that quantiles must be recognized as standardized

Table 5.2.3 *Computation, Median Family Size, Indiana, 1950*

SIZE OF FAMILY	f	PER CENT	CUMULATED PERCENTAGES
1	64,045	7.6%	7.6
2	222,047	26.4	34.0
3	185,340	22.0	56.0
4	146,519	17.5	
5	95,757	11.4	
6	57,914	6.9	
7	32,916	3.9	
8	18,500	2.1	
9	10,285	1.2	
10	5,219	.6	
11	2,469	.2	
12	1,855	.2	
	842,000	100.0%	

$$\begin{aligned} Md &= 2.5 + \left(\frac{50-34}{22} \times 1\right) \\ &= 2.5 + .73 \\ &= 3.23 \text{ persons} \end{aligned}$$

value as the median? The simple answer is that we follow the latter alternative, as we do with every other average. We do this to make finer distinctions than would be possible with whole numbers.

If, for example, the 48 states in 1950 were to be ranked by median size of family, it is obvious that fractional values would have to be employed. Our only option is to treat the data as continuous in order to satisfy the requirements of the problem which has been set up. Operating on this practical principle, we find the median of the above tabulation to be 3.23, a result which presumably would serve to establish Indiana's exact rank order in American family fertility.

The median is a simple concept: 50 per cent of the items are smaller in value, and 50 per cent are larger. Furthermore, by at least one criterion, it is the most representative of the averages: the aggregate distance between the median and each of the values is less than from any other point. It is therefore nearer to its companion values than any other average. It is in this sense that the median occupies the most central position in a distribution.

Other Position Measures. Instead of reporting a student in the upper half of his class, which requires the median, we may wish for greater precision to locate him in a smaller interval, such as the highest quarter, tenth, or

Table 5.2.4 *Yearly Family Income, White and Non-White, Urban U.S., 1955*

Income	Total	White	Non-White
Under $500	3.4%	3.0%	7.6%
$ 500– 999	4.3	3.6	11.4
1,000– 1,499	4.9	4.2	11.9
1,500– 1,999	4.9	4.5	8.9
2,000– 2,499	5.5	5.1	9.5
2,500– 2,999	5.5	5.3	8.1
3,000– 3,499	7.4	7.1	10.2
3,500– 3,999	7.2	7.2	7.0
4,000– 4,499	8.2	8.4	6.4
4,500– 4,999	7.3	7.6	4.7
5,000– 5,999	12.8	13.4	5.8
6,000– 6,999	9.5	9.9	4.8
7,000– 9,999	12.9	13.9	3.1
10,000– 14,999	4.8	5.3	0.6
15,000– 24,999	0.9	1.0	...
25,000 and over	0.5	0.5	...
Per Cent	100.0%	100.0%	100.0%
N (in '000)	42,843	38,940	3,903

Source: U.S. Bureau of the Census, *Current Population Reports, Consumer Income*, Series P-60, No. 24, April 1957, U.S. Government Printing Office, Washington, D.C., 1957.

9. For the following set of values, calculate the midpoint of the range and the median, respectively: 2, 5, 12, 16, 40. Distinguish between these two concepts.

10. In computing the median of discrete data, why is it necessary to treat them as continuous?

Section Three

The Mean, or Arithmetic Average

The Concept of Mean. Everyone has had frequent occasion to add a series of figures and divide the sum by the number of items. This is an operational definition of the mean — often used synonymously for "average" — which does not differ from the basic statistical procedure, as evidenced in the formula:

$$\bar{X} = \frac{\Sigma X}{N} \qquad (5.3.1)$$

measures of position, independent of the metric system used or of the substantive type of the data. Thus, a person, who is at the 90th centile in intelligence may also be around the 90th centile in income, thus suggesting an affinity between these two social phenomena. By such a transformation, non-comparable measures may be usefully juxtaposed.

As will he more clearly seen in a later section, these measures are also independent of the pattern of distribution — whether normal, skewed or rectangular. This circumstance enhances the versatility of quantile measures, and enables them to represent any variate as a rank position in a set of data, hy means of the statistical procedures here set forth.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Median
   Position Average
   Rank Order
   Median Rank
   Median Point
   Median Class Interval
   Decile

2. Is it possible to compute the median of a combined set of values when only the medians and total frequencies of each group are known? When total frequencies are identical?

3. Explain why the median can be calculated, even when the frequency table is open-ended.

4. In computing the median of grouped data, what assumption is made concerning the distribution of items within the median interval? Compare this with the assumption of the refined mode.

5. Is it necessary to compute the median for grouped data when one of the cumulated frequencies is equal to $\frac{N}{2}$? Verify your answer with a simple illustrative calculation.

6. (a) Compute the median of all family incomes (Table 5.2.4); of all white and non-white incomes, respectively, by Formula 5.2.1.
   (b) Could you have obtained the median of all incomes if only the distribution of incomes below $5,000 were given?
   (c) Compute quartiles for both white and non-white distributions and compare.
   (d) Would any of the above measures be affected if absolute frequencies were used instead of percentages?

7. Compute the median of the distribution of suicide rates (Table 3.1.1d), by Formula 5 2.1. Compare with the mode of the same distribution and interpret the difference.

8. Compute the median of the distribution of delinquency rates (Problem 11, p. 47). Compare with the mode.

Table 5.3.1a

Computation of Mean, Ungrouped Data

X	$X - \bar{X}$	x
0	0 − 5	−5
5	5 − 5	0
3	3 − 5	−2
9	9 − 5	4
8	8 − 5	3
25		0
where x = deviation from mean		
$\bar{X} = \frac{\Sigma X}{N} = \frac{25}{5} = 5$		

Table 5.3.1b

Computation of Mean, Grouped Data

X	f	fX	$X - \bar{X}$	x	fx
5	2	10	5 − 8	−3	−6
8	5	40	8 − 8	0	0
10	3	30	10 − 8	2	6
	10	80			0
$\bar{X} = \frac{\Sigma fX}{N} = \frac{80}{10} = 8$					

are *weighted* by their respective frequencies (f), summed, and then divided by N. The deviations (x) must likewise be weighted by class frequencies (f), and will then sum to zero.

The mean, as that point around which the positive and negative deviations exactly balance, may be illustrated graphically by a sketch of fulcrum, lever, and weights. In conformity to childhood experience on the seesaw, the aggregate force of the children on either side of the fulcrum is determined not only by their number, but also by their respective distances from the fulcrum. In the first sketch, the sum of the two weights (−5 and −2) below the fulcrum balances the two weights (+3 and +4) above the mean. In the second drawing, two weights of −3 equal the three weights of +2, as any child on a seesaw would intuitively appreciate.

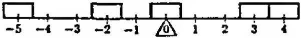


Figure 5.3.1a *Diagram of Mean as Center of Gravity, Ungrouped Data (Table 5.3.1a)*

where $\bar{X}$ = the mean (read: "X-bar")
Σ = the sum of the variates (read: "summation X")
X = a variate, or value
N = total frequency, or number of items

This is a very ancient conception of the average, and is consistent with the etymological derivation of the term. The Latin *havaria* once referred to the insured cargo loss which was equally divided among the participating shippers. This actuarial principle of spreading the risks survives, of course, as the basic procedure in mutual insurance practice. Although this arithmetic conception of the mean is quite adequate for most workaday purposes, it does not exhaust all of its statistical implications.

Another approach to the definition of the mean, necessarily congruent with the foregoing, views the values in the distribution as deviations from a central norm which is considered the true value. The scatter of shots around the bull's-eye of a target and the bell-shaped distributions of replicated laboratory measurements illustrate this conception. The deviations on both sides of the central value constitute departures, or errors, from that norm. From this point of view, the mean is considered the true value, free of errors. (See pp. 21-22).

In line with this trend of thought, deviations often tend to be discounted as being random, temporary, exceptional, or less valid than the central value. The despondent person who has encountered a "streak of bad luck" confidently expects "things to average out." He feels that something akin to a law of nature provides that bad luck in the long run must be balanced by good luck. Even a man who has enjoyed a run of good luck does not dare to "push his luck too far." The average monthly income of a salesman, the average of irregularly spaced rainfall, the batting average which is flanked by slumps and sprees, the "picnic finance" where expenses are equally divided — all these are firmly rooted in daily experience and in popular parlance. Statistically speaking, the true value emerges when the deviations cancel out

The question is, therefore: what value would we have if the errors were eliminated? The answer is: that central value around which the errors exactly balance one another, so that the positive and negative errors, when added together algebraically, equal zero. If all deviations were eliminated, it must follow that all values would be identical. The mean may therefore be defined as *that value in a given aggregate which would occur if all the values were equal.* Hence, the mean taken N times would equal the sum of the observed values: $N\bar{X} = \Sigma X$.

The basic principle of the *balanced deviations* is *computationally* illustrated in Table 5.3.1.

With ungrouped data, the mean is computed by summing the single values and dividing by N. In grouped data, however, the variates (X)

Table 5.3.1a

Computation of Mean, Ungrouped Data

X	$X - \bar{X}$	x
0	0 − 5	−5
5	5 − 5	0
3	3 − 5	−2
9	9 − 5	4
8	8 − 5	3
25		0

where x = deviation from mean

$$\bar{X} = \frac{\Sigma X}{N} = \frac{25}{5} = 5$$

Table 5.3.1b

Computation of Mean, Grouped Data

X	f	fX	$X - \bar{X}$	x	fx
5	2	10	5 − 8	−3	−6
8	5	40	8 − 8	0	0
10	3	30	10 − 8	2	6
	10	80			0

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{80}{10} = 8$$

are *weighted* by their respective frequencies (f), summed, and then divided by N. The deviations (x) must likewise be weighted by class frequencies (f), and will then sum to zero.

The mean, as that point around which the positive and negative deviations exactly balance, may be illustrated graphically by a sketch of fulcrum, lever, and weights. In conformity to childhood experience on the seesaw, the aggregate force of the children on either side of the fulcrum is

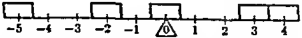


Figure 5.3.1a *Diagram of Mean as Center of Gravity, Ungrouped Data (Table 5.3.1a)*

where $\bar{X}$ = the mean (read: "X-bar")
Σ = the sum of the variates (read: "summation X")
X = a variate, or value
N = total frequency, or number of items

This is a very ancient conception of the average, and is consistent with the etymological derivation of the term. The Latin *havaria* once referred to the insured cargo loss which was equally divided among the participating shippers. This actuarial principle of spreading the risks survives, of course, as the basic procedure in mutual insurance practice. Although this arithmetic conception of the mean is quite adequate for most workaday purposes, it does not exhaust all of its statistical implications.

Another approach to the definition of the mean, necessarily congruent with the foregoing, views the values in the distribution as deviations from a central norm which is considered the true value. The scatter of shots around the bull's-eye of a target and the bell-shaped distributions of replicated laboratory measurements illustrate this conception. The deviations on both sides of the central value constitute departures, or errors, from that norm. From this point of view, the mean is considered the true value, free of errors. (See pp. 21-22).

In line with this trend of thought, deviations often tend to be discounted as being random, temporary, exceptional, or less valid than the central value. The despondent person who has encountered a "streak of bad luck" confidently expects "things to average out." He feels that something akin to a law of nature provides that bad luck in the long run must be balanced by good luck. Even a man who has enjoyed a run of good luck does not dare to "push his luck too far." The average monthly income of a salesman, the average of irregularly spaced rainfall, the batting average which is flanked by slumps and sprees, the "picnic finance" where expenses are equally divided — all these are firmly rooted in daily experience and in popular parlance. Statistically speaking, the true value emerges when the deviations cancel out.

The question is, therefore: what value would we have if the errors were eliminated? The answer is: that central value around which the errors exactly balance one another, so that the positive and negative errors, when added together algebraically, equal zero. If all deviations were eliminated, it must follow that all values would be identical. The mean may therefore be defined as *that value in a given aggregate which would occur if all the values were equal.* Hence, the mean taken N times would equal the sum of the observed values: $N\bar{X} = \Sigma X$.

The basic principle of the balanced deviations is computationally illustrated in Table 5.3.1.

With ungrouped data, the mean is computed by summing the single values and dividing by N. In grouped data, however, the variates (X)

In Table 5.3.2a, each variate (size of family) is considered to be at the midpoint of an interval of one (e.g., .5–1.5, 1.5–2.5, and so on). These variates are readily weighted by the frequencies since they are already tabulated as midpoints instead of class intervals. When, however, the class interval is larger than one, and the class limits are set down, the midpoint must first be calculated. Each midpoint is then weighted as previously, and the mean computed (Table 5.3.2b).

Table 5.3.2b *Computation of Mean Family Size, (Table 5.3.2a Regrouped)*

Family Size	Midpoint (X)	f	fX
1– 2	1.5	44	66
3– 4	3.5	32	112
5– 6	5.5	14	77
7– 8	7.5	7	52.5
9–10	9.5	3	28.5
		$N = 100$	336.0

$$X = \frac{\Sigma fX}{N} = \frac{336}{100} = 3.36$$

In allowing the midpoint to represent all the values in a given class interval, we resort to the assumption that all the items in the interval have the value of the midpoint. Even though this is not true, the assumption is not too violent, and is in keeping with the good statistical practice of accepting approximation as a small price for expediting calculations.

The mean of the regrouped distribution (3.36) differs slightly from that of the previous calculation (3.35), a discrepancy which is merely a result of grouping error.

The Concept of Coding. Since, in actual practice, frequencies as well as magnitudes are likely to be large, the foregoing methods will usually become quite cumbersome. Hence, certain arithmetic tactics have been devised to simplify the pencil-and-paper computatious and to reduce the possibility of errors. One type of operation has come to be known as *coding*.

Coding is a generic term which may denote: (1) the substitution of a convenient simplified symbol for a set of complex values, or (2) the reduction of a mass of data into simpler sets of categories. Consequently, the general concept has wide applicability. The first type is exemplified by the code employed in the preparation of data for electronic sorters and

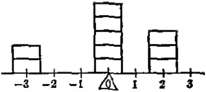


FIGURE 5.3.1b *Diagram of Mean as Center of Gravity, Grouped Data (Table 5.3.1b)*

Calculation: Grouped Data. As the student now knows, grouped data do not differ from simple ungrouped data except in the fact that similar values are grouped together, and the frequency, or *weight*, of the grouped items is indicated for each group in the "*f*" column. Therefore, to calculate the mean of grouped items, we obtain the sum of the weighted values, and divide by *N*.

Table 5.3.2a

Computation of Mean Family Size, One Hundred Families

X	f	fX
1	20	20
2	24	48
3	17	51
4	15	60
5	9	45
6	5	30
7	4	28
8	3	24
9	1	9
10	2	20
	$N = 100$	335

$$\bar{X} = \frac{\Sigma fX}{N}$$
$$= \frac{335}{100}$$
$$= 3.35 \text{ persons per family}$$

Source: Hypothetical

To determine, for example, the mean number of persons per family, we must first determine the total number of persons (ΣfX), and then divide by the number (N) of families. Since there are 20 *families of one person* each, the weighted total in this class will be 20; 24 families of 2 persons each give a total of 48 persons. Proceeding as in Table 5.3.2a, we find the mean to be 3.35 persons per family.

Although this average, based as it is on discrete data, cannot correspond to any observed size, it is nevertheless a useful measure of magnitude, which, as with the median, no one finds disturbing.

Table 5.3.4

Computation of Mean, Coding by Subtraction

Original Values	Coded Values
5,000,003	3
5,000,007	7
5,000,008	8
15,000,018	18
$\bar{X} = 5{,}000{,}006$	$\bar{X} = 6$
$\bar{X} = 5{,}000{,}000 + 6$ $= 5{,}000{,}006$	

remainders; and finally (3) to bring forth the 5,000,000 and restore it to the coded mean. We thereby obtain the mean of the original values. Any other conceivable *coding constant* could have been subtracted, and the final result would have been the same. But if the objective is maximum simplicity, the only practical choice would be 5,000,000 as our coding constant.

Method of the Guessed Mean. The foregoing procedure reduces itself to the method of the *guessed mean.*[*] By inspection of the oversimplified example quoted above, it is clear that the mean would be close to 5,000,000. We could have accepted this figure as a tentative mean, and viewed the coded variates as deviations from the preliminary guess. We would then "correct" the guessed mean by an amount equal to the mean of the deviations from the guessed mean. The full procedure would be carried out according to the following formula:

$$\bar{X} = \bar{X}' + C \qquad (5.3.2)$$

where $\bar{X}$ = the mean

$\bar{X}'$ = the guessed mean

$X - \bar{X}'$ = deviation from the guessed mean

$C = \dfrac{\Sigma(X - \bar{X}')}{N}$, or the *correction factor*

For the data in Table 5.3.5, we set the guessed mean equal to 5, and compute the deviations about this provisional mean. It could hardly ever be expected that the guessed mean would turn out to be the true mean. But, from the definition of the mean, we know that if the choice had been correct, the sum of the deviations would have been zero, and the problem would be solved. Since the mean of the deviations exceeds zero by 1.4, we must raise the guessed mean by that amount, which correspondingly depresses

[*] This method of the guessed mean is also known by other labels, consistent with this designation: *assumed mean, provisional mean,* and *arbitrary origin.*

computers, or the Morse code in telegraphy. The second type is illustrated by the arithmetic simplification of bulky figures with which we are here concerned. However, unlike the codes which a military enemy is determined to decipher or "break," complexity in a statistical code is not considered a protection or a virtue, but rather a defect.

The type of coding here under consideration rests on the elementary *principle* that *uniform* arithmetic treatment of a series of values leaves corresponding relations unaltered: subtraction or addition of constants leaves intervals between values unchanged; multiplication or division by like factors leaves ratios between values within the series unaltered. This is demonstrated in the series shown in Table 5.3.3: the subtraction of the constant 50 from each value does not alter differences between values. Analogously, the division of each value by 50 does not alter ratios among values. Thus, 200:100 :: 4:2, and 1,000:100 :: 20:2.

Table 5.3.3

Subtracting a Constant

X	D	$X - 50$
100		50
	100	
200		150
	100	
300		250
	200	
500		450
	500	
1,000		950

In statistics such coding is designed only as a temporary convenience. Consequently, after the coded computations have been completed, the results must be decoded: the constant factor used in division must be reintroduced by multiplication; and, the factor subtracted must be added in order to restore the original quantities, and thereby produce the results which would have been obtained without encoding in the first place. By adding 50 where we had previously subtracted it, or by multiplying by 50 where we had divided by it, it is obvious that we have merely re-established the original values as though nothing had happened.

Coding Applied to the Mean. Let us suppose that we are required to find the mean of the three large numbers shown in Table 5.3.4. The given values sum to 15,000,018 and have a mean of 5,000,006. Now it will be evident that it was wasteful to carry along the heavy freight of 5,000,000 It would have been much simpler: (1) to subtract the 5,000,000 from each value, to sequester it, as it were; (2) to find the mean of the manageable

Table 5.3.4

Computation of Mean, Coding by Subtraction

ORIGINAL VALUES	CODED VALUES
5,000,003	3
5,000,007	7
5,000,008	8
15,000,018	18
$\bar{X} = 5{,}000{,}006$	$\bar{X} = 6$
$\bar{X} = 5{,}000{,}000 + 6$ $= 5{,}000{,}006$	

remainders; and finally (3) to bring forth the 5,000,000 and restore it to the coded mean. We thereby obtain the mean of the original values. Any other conceivable *coding constant* could have been subtracted, and the final result would have been the same. But if the objective is maximum simplicity, the only practical choice would be 5,000,000 as our coding constant.

Method of the Guessed Mean. The foregoing procedure reduces itself to the method of the *guessed mean.** By inspection of the oversimplified example quoted above, it is clear that the mean would be close to 5,000,000. We could have accepted this figure as a tentative mean, and viewed the coded variates as deviations from the preliminary guess. We would then "correct" the guessed mean by an amount equal to the mean of the deviations from the guessed mean. The full procedure would be carried out according to the following formula:

$$X = X' + C \qquad (5.3.2)$$

where X = the mean

X' = the guessed mean

$X - X'$ = deviation from the guessed mean

$C = \dfrac{\Sigma(X - X')}{N}$, or the *correction factor*

For the data in Table 5.3.5, we set the guessed mean equal to 5, and compute the deviations about this provisional mean. It could hardly ever be expected that the guessed mean would turn out to be the true mean. But, from the definition of the mean, we know that if the choice had been correct, the sum of the deviations would have been zero, and the problem would be solved. Since the mean of the deviations exceeds zero by 1.4, we must raise the guessed mean by that amount, which correspondingly depresses

* This method of the guessed mean is also known by other labels, consistent with this designation: *assumed mean, provisional mean,* and *arbitrary origin.*

computers, or the Morse code in telegraphy. The second type is illustrated by the arithmetic simplification of bulky figures with which we are here concerned. However, unlike the codes which a military enemy is determined to decipher or "break," complexity in a statistical code is not considered a protection or a virtue, but rather a defect.

The type of coding here under consideration rests on the elementary principle that uniform arithmetic treatment of a series of values leaves corresponding relations unaltered: subtraction or addition of constants leaves intervals between values unchanged; multiplication or division by like factors leaves ratios between values within the series unaltered. This is demonstrated in the series shown in Table 5.3.3: the subtraction of the constant 50 from each value does not alter differences between values. Analogously, the division of each value by 50 does not alter ratios among values. Thus, 200:100 :: 4:2, and 1,000:100 :: 20:2.

Table 5.3.3

Subtracting a Constant

X	D	$X - 50$
100		50
	100	
200		150
	100	
300		250
	200	
500		450
	500	
1,000		950

In statistics such coding is designed only as a temporary convenience. Consequently, after the coded computations have been completed, the results must be decoded: the constant factor used in division must be reintroduced by multiplication; and, the factor subtracted must be added in order to restore the original quantities, and thereby produce the results which would have been obtained without encoding in the first place. By adding 50 where we had previously subtracted it, or by multiplying by 50 where we had divided by it, it is obvious that we have merely re-established the original values as though nothing had happened.

Coding Applied to the Mean. Let us suppose that we are required to find the mean of the three large numbers shown in Table 5.3.4. The given values sum to 15,000,018 and have a mean of 5,000,006. Now it will be evident that it was wasteful to carry along the heavy freight of 5,000,000 It would have been much simpler: (1) to subtract the 5,000,000 from each value, to sequester it, as it were; (2) to find the mean of the manageable rates. (1) The midpoint of the interval within which the true mean probably lies is selected as the guessed mean. This is usually, but not necessarily, the class interval of the largest frequency. We choose 13. (2) The guessed mean is then subtracted from each midpoint (Column 4); that is, it is used as a subtraction code. The deviations resulting from this operation — 3, 6, 9, and so on — are all necessarily multiples of the class width, 3. This suggests (3) *coding by division*, using the class interval as the divisor. The results of this operation are designated x' (Column 5). (4) We now weight the twice-coded midpoints by their respective frequencies (Column 6) and compute their mean: $\frac{\Sigma fx'}{N} = \frac{-18}{107} = -.168$. This completes the encoding operation and sets the stage for the reverse course of *decoding*. Accordingly, (5) we multiply the double-coded mean by class width and thereby obtain the correction factor:

$$\begin{aligned} C &= \left(\frac{-18}{107} \times 3\right) \\ &= -.50 \end{aligned}$$

We finally (6) add this correction factor to the guessed mean in order to find the true mean:

$$\begin{aligned} \bar{X} &= 13 + (-.50) \\ &= 12.5 \end{aligned}$$

Putting the separate steps together we obtain the complete working procedure:

$$\bar{X} = \bar{X}' + \left(\frac{\Sigma fx'}{N} \times i\right) \qquad (5.3.3)$$

Substituting the quantities calculated above, we have

$$\begin{aligned} \bar{X} &= 13 + \left(\frac{-18}{107} \times 3\right) \\ &= 13 + (-.50) \\ &= 12.5 \end{aligned}$$

Mechanical Routine. The foregoing procedure may at first appear circuitous and complicated. However, it becomes very simple in actual operation, since the twice-coded values may be more quickly obtained without the intervening step presented in Column 4. Instead, we simply designate as zero the interval of the guessed mean, from which we measure the interval-deviations, as in Column 5.

This abbreviated routine is illustrated in Table 5.3.6b, which is essentially a worksheet on which the necessary steps are carried out. It yields all of the quantities required by the formula.

The manner of calculation demonstrates effectively the economy of selecting the guessed mean in the general neighborhood of the true mean.

each deviation around the guessed mean by the same amount. The fact that the correction factor is positive indicates that the guessed mean was too low.

Table 5.3.5

Computing the Mean, Method of the Guessed Mean, $\bar{X}' = 5$

X	$X - \bar{X}'$
2	−3
5	0
7	2
8	3
10	5
32	7
$\bar{X} = 6.4$	$C = \frac{\Sigma(X - \bar{X}')}{N}$ $= 1.4$
$\bar{X} = \bar{X}' + C$ $= 5 + 1.4$ $= 6.4$	

Grouped Data. Of course, the above trivial illustration does not require the short-cut technique of the guessed mean. It is grouped data which stand in need of such assistance. And here it is expeditious to code not only by subtraction, but by division as well. We will demonstrate the steps in this procedure (Table 5.3.6a) on the frequency table of 107 suicide

Table 5.3.6a *Computation of Mean, Grouped Data, Suicide Rates, 1950*

(1) RATE	(2) MDPT. (X)	(3) f	(4) $X - \bar{X}'$	(5) $\frac{X - \bar{X}'}{i} = x'$	(6) fx'
3– 5	4	6	−9	−3	−18
6– 8	7	18	−6	−2	−36
9–11	10	29	−3	−1	−29
12–14	13 ($\bar{X}'$)	24	0	0	0
15–17	16	13	3	1	13
18–20	19	7	6	2	14
21–23	22	4	9	3	12
24–26	25	4	12	4	16
27–29	28	2	15	5	10
		$N = 107$			−18

It is ohvious, however, that in many problems, unequal weights will be encountered. In such a case, the means of the suhgroups, prior to their being combined, would have to be specifically weighted hy their respective N's.

Thus, if the two housing units, with grade-point averages of 1.4 and 1.8 had populations of 26 and 42 respectively, each mean would have to he accorded its proper weight in order to obtain the mean of the comhined group. This is done hy multiplication, which constitutes the procedure of weighting (Tahle 5.3.7). The resemblance to the familiar frequency table is obvious. In fact, the summing of the ordinary frequency table is a process of weighting each midpoint by its frequency.

Table 5.3.7

Computation of Weighted Mean

$\bar{X}$	f	$f\bar{X}$
1.4	26	36.4
1.8	42	75.6
	68	112.0

$$\text{Weighted } \bar{X} = \frac{112}{68}$$
$$= 1.65 \text{ grade points per student}$$

The failure to weight often leads to absurd results. We cite an illustration from baseball:

> A batter, in 75 trips to the plate, has amassed 25 hits for a batting average (mean) of .333; on this day he gets 5 hits in 5 trips, for a day's average of 1,000. What is the combined average?

A naive grandstand "statistician" may add the two averages, and then divide by 2, for a comhined average of .667 — a patently unreasonable result. The proper calculation is shown in Table 5.3.8.

Table 5.3.8

Mean of Combined Sets, Unequal N's

$\bar{X}$	f	$f\bar{X}$
.333	75	25
1.000	5	5
	80	30

$$\text{Weighted } \bar{X} = \frac{30}{80}$$
$$= .375 \text{ (batting average)}$$

Table 5.3.6b *Worksheet for Computation of Mean, Grouped Data*

(1) RATE	(3) f	(5) x'	(6) fx'
3– 5	6	−3	−18
6– 8	18	−2	−36
9–11	29	−1	−29
12–14	24	0	0
15–17	13	1	13
18–20	7	2	14
21–23	4	3	12
24–26	4	4	16
27–29	2	5	10
	$N = 107$		−18

However, the principles of arithmetic do not require this selection. Any other "guessed" value will do the work — but less efficiently. In machine calculation, however, when arithmetic economy is not such an important factor, the calculation may be simplified by arbitrarily setting the guessed mean at zero, and coding from zero — which is equivalent to using the given raw values.

The Weighting of Means. The *weight* of a value, statistically speaking, is simply its frequency. Any collective measure may, therefore, be considered as having a weight equal to the number of observations it represents. The mean is such a collective measure: it is not itself an observed value; it is a derived figure. Consequently, like ratios and percentages, which are also derived values, the mean should never be manipulated without taking its weight into consideration.

Two or more means are, themselves, frequently averaged; that is, we may take the mean of a series of means. If the means derive from equal N's, i.e., are of equal weight, there is no special difficulty in the operation. Consider the following problem:

> A group of 42 students, living in one housing unit, have a grade-point average of 1.4, another group of 42 students, in another housing unit, have a grade point average of 1.8. What is the average for the 84 students?

Since the grade-point means have equal weight, one need only add the two means and divide by 2:

$$\bar{X} = \frac{1.4 + 1.8}{2}$$

$$= 1.6 \text{ grade points per student}$$

operations, as in the preparation of budget norms used by welfare workers, or cost of living indexes employed by economists.

Furthermore, as with many other prevalent statistical concepts, the term possesses fringe meanings which may at first seem to he inconsistent with the quantitative meaning here assigned. Some of these meanings are more metaphorical than material — as, for example, the weight assigned to one person's opinion or the weight of an argument. When more closely inspected, however, the varieties of connotations will be found to have a pervasive common core which leaves the utility of the concept unmarred.

Utility of the Mean. Although the mean is the most widely used average, it is not of universal applicahility. Its fundamental characteristic, namely, that it reflects the magnitude of every item in the array, renders it at the same time hoth appropriate and inappropriate for specific purposes. Since it reflects every value in the array, it will be affected hy the extremely high or low values that are always found in a skewed distribution, and it will therefore lose its typicality and perhaps mislead the reader. This phenomenon had led some to assert that the mean should not be used when the data are markedly skewed.

But such a conclusion confuses the arithmetic and the conceptual aspects of this statistical calculation. In a statistical sense, it is true, the mean may not he typical; however, in a substantive sense, the mean value may be the very one desired. In the case of "picnic finance," for example, it is intended that the extreme magnitudes (e.g., a person's outlay for expensive meat) be reflected in the equal division of expenditures. On the other hand, if a young lawyer were estimating the size of his prospective income, he would not compute an arithmetic average of the combined incomes of corporation attorneys and those of the less prosperous but more numerous solo lawyers. The mode would furnish a better indication of his prospects. The inappropriateness of the mean for this set of data lies in the heterogeneity of the lawyer group, in which the young as well as the experienced are indiscriminately merged.

The fact that the magnitude of every item is accurately weighted endows the mean with a certain mathematical character which is not shared hy any other average. Hence, the mean may be used to huild up subsequent computations. We have demonstrated that means can, themselves, be averaged and weighted, can be segmented and recombined, a quality which is not possessed hy the mode and the median. Means can be manipulated algebraically because they have been obtained algehraically. This extraordinary utility of the mean confers upon it a prestige that sometimes undeservedly overshadows the more limited but still quite specific and indispensable qualities of the other averages.

Unweighted Means. The logical requirement of weighting is not always as clear cut as it is in the above illustration of the batting averages. There are occasional instances in which discretion may be employed concerning whether to weight in the conventional manner as described above, or whether to disregard the *N*'s from which they have been derived.

Let us suppose that we wish to calculate the grade average of a large number of housing units. The previously described procedure of weighting by the number of students in each housing unit would naturally yield the grade-point average *per student.* However, if we desire the grade average *per housing unit,* we would ignore the size of the population of each housing unit and compute the average of the house averages. Such a mean is usually labeled an *unweighted average.*

In a certain sense, of course, there is no such thing as an unweighted average; it is only a question of which weight we select, and this is determined by the nature of the information desired. To put it concisely, it is a question of the average "per what" we demand. In the following problem, the choice of the unweighted average will again seem most reasonable.

Ten of the largest cities in the United States have suicide rates (per 100,000 population) as follows: 5, 9, 3, 0, 15, 6, 10, 2, 1, 7. What is the mean?

Presumably, we could weight the rate of each city according to its population size, and compute the combined rate, as we did for the batting average. However, we are not likely to be interested in the suicide rate of such a population aggregate, since this aggregate of persons in the ten individual cities does not constitute a meaningful social unit. Rather, we are more likely to be interested in the average rate *per city* in this category; hence, each city would be treated as a unit, and not weighted for *population size. This average is an unweighted* mean *of the rates for the* individual cities:

$$\begin{aligned} \bar{X} &= \frac{58}{10} \\ &= 5.8, \text{ the mean rate per city} \end{aligned}$$

There can, of course, be no routine, ironclad rule on the issue of weighting. It is a question of purposes and consequences Consider, for example, the consequences of weighting the above list of cities by population, of which New York City is one. That city would then overwhelm the rates of all other cities and produce a composite rate that would differ very little from that of the great metropolis. For this reason, as well as the even more pertinent one that the unit of interest is the individual city and not the person, the unweighted average is called for.

Although the concept of weighting is here applied only to means, rates, and percentages, it is widely applicable to many complicated combining

12. In Table 5.3.9, if the letter grades are scored as follows: A (3), B (2), C (1), what is the student's grade point per course? per hour?
13. Compute the mean of the frequency distribution of delinquency rates (Problem 11, p. 47), according to Formula 5.3.3.

Section Four

Criteria for Choice of Average

Because we are frequently faced with alternative choices among the various averages, it is necessary to lay down a few principles which will serve as guides in the selection of a central value. To a certain extent, this problem has already been clarified. It has been stated that there is no all-purpose average which can be universally employed. It should always be kept in mind that an average is a single representative value which possesses the convenience of compactness, but also the inconvenience of its brevity. Even at best, it will conceal and exclude as much or more information than it reveals of the distribution from which it is extracted.

Three major criteria, in order of priority, which may be used in judging the applicability of an average are: (1) the purpose to be served, or the question that it is designed to answer; (2) the pattern of the distribution of the data; (3) various technical considerations, primarily of an arithmetic nature, which limit the choice of average.

Purpose To Be Served. Any empirical sociological inquiry is essentially an attempt to answer a question of a substantive nature, for which the statistical procedure is only a tool. Questions to which one or another average would supply the answer could be: What is the size of the American family? What is the length of life? What is the age of the American population? Such questions must be further interpreted in terms of the purposes of the inquiry. If the average size of the American family is desired for purposes of planning a housing development, the crude mode would be more pertinent than the mean or median, even though the precise degree of modality is unknown. Houses are not built for non-existent arithmetic averages, but for actual families that do exist with a certain frequency. If on the other hand, the average is to serve as a measure of general fertility, the mean would be more relevant, since it would reflect the contribution of the most fertile as well as the least fertile. For actuarial purposes, the equal distribution of risks by very definition requires the mean length of life.

Pattern of Distribution. Distributions may, of course, range in degree from *symmetrical* to extremely *skewed.* Symmetry means that the values are

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Arithmetic Mean
   Coding
   Weighting
   Weighted Mean
   Guessed Mean

2. Explain why the sum of the deviations around the mean is zero.

3. In computing the mean of grouped data, what value is assigned to all items in a given class interval? Is this a reasonable procedure? Explain.

4. Give several instances when the arithmetic mean would be appropriate, even though the scatter of items is highly skewed.

5. When zero is used as the guessed mean in computing the mean, what is the value of C, or the correction factor?

6. Discuss the validity of the statement that the extreme values in a skewed distribution make a disproportionate contribution to the arithmetic mean.

7. Outline the procedure for finding the mean of two or more groups when only the total frequency and the arithmetic mean of each group is given. Express this procedure symbolically.

8. If the mean of 100 items is 10, and the mean of 50 items is 15, whet is the meen of the combined group? What would the mean be if each group had 50 items? 100 items?

9. A group of 25 families hes a mean weekly income of $45.00; a second group of 15 femilies has a mean weekly income of $60.00. What is the *total* income of the combined group?

10. If the mean of a series is 6 and the guessed mean is 5, by how much on the average is each deviation from the guessed mean too large?

11. A student received letter and percentage grades in his courses as shown in Table 5 3.9. What is his percentage grade per course? per hour? Which average would be used? Would you weight one or both averages? Explain.

Table 5.3.9 *Student's Grade Record*

SUBJECT	CREDIT HOURS	GRADE	
		Percentage	*Letter*
History	5	92	A
English Literature ...	3	85	B
Psychology Lab......	2	72	C
French	5	95	A

however, additional computations are anticipated, the mean and its derivatives must be selected.

Minimum, Maximum, and Intermediate Measures. In many statistical discussions, only the averages seem to be considered as possible measures of location. However, it should be re-emphasized that this is a limited conception. Measures of location are not necessarily measures of typicality. Averages tend to connote typicality or representativeness; but the maximum, minimum or any intermediate value may serve to fix location, provided it is selected from the whole array. For example, although many a prospective university student (or his parent) may find the average expenditure on the campus a useful guide, an impecunious student may be interested only in the minimum, and an affluent one only in the maximum, as descriptive of the standard of living on that campus. Consequently, it should be reiterated that an average must be viewed as only one of the many measures of location, each appropriate to a specific problem.

Characteristics of Averages. The selection of the proper average presents problems not posed by other measures of location. Hence, a thorough knowledge of the descriptive characteristics of each average is a prerequisite to their proper employment. The characteristics of the different averages are frequently discussed in terms of their "advantages and disadvantages." But these terms are evaluative, rather than descriptive, and therefore have no fixed meaning. An advantage in one context may be a disadvantage in another. Hence, we prefer to set forth the characteristics which are intrinsic in the averages, quite independent of the setting in which they may be used.

Summary of Characteristics of Averages

The Mode

1. It is the *most frequent value in the distribution; it is the point of* greatest density.
2. The value of the mode is established by the predominant frequency, and not by the values in the distribution.
3. It is the most probable value, and hence the most typical.
4. A given distribution may have two or more modes. On the other hand, there is no mode in a rectangular distribution.
5. The mode does not reflect the degree of modality.
6. It cannot be manipulated algebraically: modes of subgroups cannot be combined.
7. It is unstable in that it is influenced by grouping procedures.
8. Values must be ordered and grouped for its computation.
9. It can be calculated when table ends are open.
10. It is the only average which can be applied to qualitative variables.

distributed identically on either side of the mean; while skew is merely the absence of symmetry. The degree of skewness affects the typicality and representativeness of the average values, and hence must be taken into account in judging their relevance.

It is sometimes alleged that, if the frequency curve is symmetrical, there is no problem of choice, since mean, median, and mode are identical. This is true, however, only in an arithmetic sense, not in a conceptual sense. *Even with the same numerical value, there is a different imagery attached* to the concepts of the mean, mode, and median. For example, the "average" student does not conceive of his grade as being equal to the sum of *all grades divided by N, but rather conceives of it as a position in the array.* The young physician, new to the community, in estimating his probable income, does not concern himself with the mean, but with the modal income, even though quantitatively they may be identical. From these illustrations we may conclude that, even though the distribution is symmetrical, we still select the average which conceptually satisfies our purposes.

As the distribution becomes more and more skewed, the values of the averages diverge correspondingly, and the choice of the average becomes more critical. In such cases, the mean loses its typicality, in that it is less likely to be empirically encountered. In the U-curve, in fact, the mean may be practically nonexistent, and therefore may be quite unrealistic. Since many sociological data — wages, sizes of cities, sizes of families, and the like — are often severely skewed, it is essential to consider the pattern of the curve in the selection of the most suitable average.

Technical Considerations. There are certain purely technical features of a tabulation which may compel the use of one or another average. Thus, since the mean cannot be calculated for open-ended data, the median may have to be resorted to. *But this is often quite satisfactory* provided the distribution is not unduly skewed, in which event mean and median do not differ much in any case.

On the other hand, the mean is the only average computable when only the total values and N are known, even though another average might have been preferred. This is illustrated by a citation from the U.S. Census, 1950:

$$
\begin{aligned}
\bar{X} &= \frac{\text{Total Population in Families}}{\text{Number of Families}} \\
&= \frac{138{,}079{,}600}{38{,}310{,}980} \\
&= 3.6 \text{ persons per family}
\end{aligned}
$$

The technical requirements of possible subsequent calculations should always be kept in mind. Since there is no method for combining or weighting medians and modes, they tend to become terminal measures. When, Being influenced by each individual value, the mean of a skewed distribution will be pulled in the direction of the extreme values. The mode, of course, is not influenced by the flanks of the distribution; and the median is drawn toward the tail solely by the relative frequency of items in that tail. However, this latter attraction is not very great because the concentration of items in the tail is necessarily sparse. Accordingly, when the skew of a unimodal distribution is to the right (Figure 5.4.1), the order of the averages on the base line will be: mode, median, and mean; and the gap between them will vary according to the severity of the skew. A skew to the left will, of course, reverse the order.

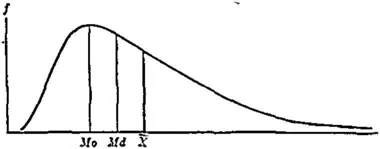


Figure 5.4.1 *Mean, Median, and Mode, Right Skew*

But even two averages of the same type cannot be safely compared unless the patterns of distribution are similar. It is not too reckless to proffer the maxim that two or more averages should never be compared unless "all other things are equal" or at least similar. The comparison of the mean height of males ($\bar{X}$) and that of females ($\bar{Y}$) (Figure 5.4.2) is

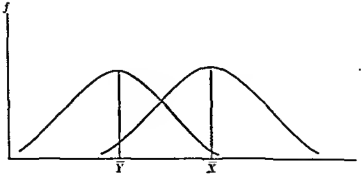


Figure 5.4.2 *Mean Heights of Males* ($\bar{X}$) *and Females* ($\bar{Y}$)

The Median

1. *It is the value of the middle point of* the array (not midpoint of range), such that half the items are above and half below it.

2. The value of the median is fixed by its position in the array, and does not reflect the individual values.

3. The aggregate distance between the median point and all the values in the array is less than from any other point.

4. Each array has one and only one median.

5. It cannot be manipulated algebraically: medians of subgroups cannot be weighted and combined.

6. It is stable in that grouping procedures do not appreciably affect it.

7. Values must be ordered, and may be grouped, for computation.

8. It can be computed when ends are open.

9. It is not applicable to qualitative data.

The Mean

1. It is the value in a given aggregate which would obtain if all the values were equal.

2. The sums of the deviations on either side of the mean are equal; hence, the algebraic sum of the deviations is equal to zero.

3. *It reflects the magnitude of every value.*

4. An array has one and only one mean.

5. Means may be manipulated algebraically: means of subgroups may be combined when properly weighted.

6. It may be calculated even when individual values are unknown, provided the sum of the values and N are known.

7. Values need not be ordered or grouped for its calculation.

8. It cannot be calculated *from* a *frequency* table when ends are open.

9. It is stable in that grouping procedures do not seriously affect it.

10. It is applicable only to quantitative data.

Comparability of Averages. Like every other statistical quantity, averages are used for comparative purposes. Specifically, averages are used to compare the locations of distinct groups or distributions on the same scale. It should be obvious, however, that the different types of averages are not comparable That is to say, the mean of one distribution cannot be legitimately compared to the mode of another in determining their relative locations on a given continuum. The reason for this prohibition is simply that the various types of averages are not coordinate measures; they reflect entirely different aspects of a distribution: the mode reflects the highest *frequency*, the mediăn reflects the middle *position*, and the mean reflects the *centrality of values.*

These now familiar differentiating characteristics become all the more apparent when the effect of *skewness* on the averages is taken into account.

approximately the same range; the means are separated, however, because of the skew in one of the distributions. Here, too, the modes, which are almost identical, would do greater justice to the location of the data. Of course, when dominance of location is less pertinent than centrality of values, our comparison would necessarily be based on means.

Conclusion. The foregoing discussion inevitably leads to the truism which could have been anticipated: averages, like every other statistic, are incomplete descriptions of a set of data. They should not be conceived of as autonomous, but rather as being bound to the data from which they are derived. A responsible use of the average — or, for that matter, any other measure of location — must take into consideration the tendency of the user to reconstruct mentally the distribution from which only a single value has been extracted. To supplement this incomplete description afforded by the average, it is essential to examine the scatter or dispersion of the distribution relative to the average. This is the subject of the next chapter.

Questions and Problems

1. Define the following concepts:
   Symmetry
   Skew
   Right (Positive) Skew
   Left (Negative) Skew
   Unimodal
   Rectangular Distribution
   Maximum
   Minimum

2. State in general terms how right skew affects the three common averages; left skew.

3. In a frequency distribution of family incomes in the United States (Table 5.2.4), how would the median, mode, and mean rank by order of magnitude?

4. As a result of the mental tests administered to American soldiers during World War I, it was reported that "most Americans were below the average score." Is that possible? Probable? Which average could have been meant?

5. When would the maximum value best represent a frequency distribution? The *minimum*?

6. How would the median and mean grades in a certain class be affected if:
   (a) the poorest students in the class withdrew?
   (b) the poorest students were successfully tutored?
   (c) the middle students were successfully tutored?
   (d) an easier examination were given?

quite appropriate. Men are taller than women, and the degree of difference is properly measured by the differences in their means. Males and females are located merely on different segments of the same scale of heights; their patterns of variation are identical.

However, when the pattern of distribution differs markedly, such comparisons may in fact be very misleading, especially in the case of means. For example, when two means are quoted as identical, the reader may be misled into reconstructing the curves imaginatively also as identical patterns. In truth, however, the two series may present quite different contours, as in Figure 5.4.3.

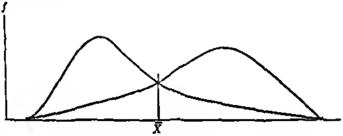


FIGURE 5.4.3 *Opposite Skews, Identical Means*

To be sure, in a gross sense, the locations of the two sets of data are identical, for they cover exactly the same range. But the locations of the predominant segments of the distributions are widely separated, and would be more accurately specified by the modes. In this instance, the identical means are a consequence of the overlapping tails, rather than a reflection of the overlapping humps.

In Figure 5 4.4, the locations of the predominant frequencies are within

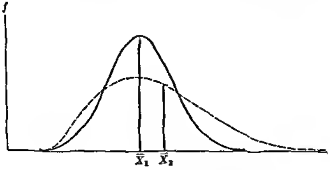


FIGURE 5.4.4 *Effect of Skew on the Mean*

Variation 6

Section One

Measures of Spread

*Concept of Variation.** Variation is the occasion for all statistics. If all the values in a given set were identical, it would be superfluous to calculate an average — or, for that matter, any other statistical measure — since any single measure would already accurately represent all. The very purpose of averaging is of course to provide a single value to represent a group of unlike values. In fact, averages were invented to suppress the differences among values whenever such differences are not pertinent to the case.

However, *under certain circumstances these differences may be of as* much or even more interest than the average itself. Thus, in arriving at a final grade, a teacher will consider not only the average of the student's test scores, but the spread of those scores as well. A student with marks of 100, 90, and 50 will be differently evaluated from one with scores of 80, 83, and 77, although both present the same mean score of 80. In selecting for varsity competition between two players with equal scoring averages, the basketball coach is more likely to use the consistent player who is seldom off that average, rather than the erratic performer who is generally low but now and again spectacularly high. Analogously, two occupational groups having approximately the same mean annual income — professors and business executives, for example — may nevertheless present very different patterns of income opportunities. In the teaching profession,

* Many textbooks do not distinguish between the terms *variability* and *variation*. We restrict ourselves to the concept variation, defining variability as the capacity to vary, and variation as the manifestation of that capacity which we endeavor to describe and measure. Without variation, there could be no statistics in the first place. For that reason, statistics is now and then referred to as the "science of variation."

7. In what type of distribution would the median student not be included among the modal students? Show graphically.

8. Does the expression, "the average Englishman" have any statistical meaning? Can the concept be quantified? Similarly: the average married man, average eye-color, average Catholic, average American, average taxpayer.

9. Identify the averages in the following hypothetical illustration: The average American father in 1958 was aged 44, had 1.5 children, lived in a town of 2,500, was a native American, spent $1,200 in retail stores.

10. Distinguish between the two statements: The average American male marries at age 26; the American male marries at the average age of 26.

11. List the absolute minimum information necessary for the calculation of the *mean*, *mode*, and *median* from grouped data.

12. In Community A, the modal length of life is 55, the median is 60, and the mean is 65; in Community B, the modal length of life is 70, the median is 65, and the mean is 60. From this information, reconstruct the frequency curves. Which community is the healthier?

SELECTED REFERENCES

McCarthy, Philip J., *Introduction to Statistical Reasoning.* McGraw-Hill Book Company, Inc., New York, 1957. Chapter 4

Moroney, M. J., *Facts from Figures.* Penguin Books Limited, Harmondsworth, Middlesex, 1954 Chapter 4.

Yule, G. Udny, and M. G. Kendall, *An Introduction to the Theory of Statistics.* Fourteenth edition. Hafner Publishing Co., New York, 1950. Chapter 4.

and below the mean are *asymmetrical* to each other; whether the values become more numerous as we approach the mean (*unimodal*), or more numerous as we approach the extremes of the range (*U-shaped*); whether the values distribute themselves uniformly over the entire range with no point of concentration (*rectangular*), or congregate at rival points along the range as in a *bimodal* distribution. Also, we may judge from the tail of the frequency curve whether the *skew* is to the right or to the left, as well as the severity of the skew. Finally, by superimposing one curve on another, it is possible to rank distributions according to their degree of *kurtosis*, a term used to describe the extent to which a unimodal curve is peaked.

However, such impressionistic judgments of variation have a restricted utility. In fact they may be quite misleading, affected as they are by the arbitrary graphic scales. Moreover, such visual impressions are personal and subjective and therefore can never be precisely communicated. They therefore serve to point up the need for objective indexes of variation which have a standard meaning.

Measurements of Range. The crudest and simplest measure of variation is the *total range*, or merely the *range*. By definition, the range is that interval which encompasses all of the values. Consequently, it is calculated in exactly the same manner as a class interval: we take the difference between the true extremes of the array, which in this case constitute the interval boundaries. Thus, to find the range of the suicide rates (Table 3.1.1c), we identify the extreme true values and then subtract one from the other. The smallest rounded rate is 3 and the largest is 29; therefore, the range is the difference between 2.5 and 29.5, or 27.

This is the minimum span required to accommodate all of the observed rates. Had a larger or smaller number of rates been recorded and included in the set, the range would undoubtedly have been otherwise. For example, the range of rates of cities with 50,000 inhabitants and over would in all likelihood have exceeded 27, while the range for the less numerous cities of 250,000 and over would probably have been smaller. By augmenting the original group of observations, the range can either expand or remain unaltered; it cannot shrink. For this reason, two or more ranges should in general not be compared unless they are based on approximately the same number of items. For example, it would be inappropriate to compare the score range of two students, unless each range embraced about the same number of test scores.

For discrete data, the procedure is the same as that for continuous data, except that the true limits are now a necessary fiction. Thus, an array of family sizes of 2 through 12 has a range of 11, which is the difference between 12.5 and 1.5. This means that the variable can take 11 and only 11 consecutive values: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12. Some texts label this the *inclusive range* and define it as $(H - L) + 1$; e.g., $(12 - 2) + 1$.

salaries are more standardized around the average with small likelihood of extreme poverty or wealth, whereas, in business, incomes are more widely dispersed, with both greater risk of poverty and occasional opportunity for wealth.

In general, therefore, the pattern of variation among the observed values — what statisticians call *spread, scatter,* or *dispersion* — is of as much relevance as is the location of the distribution. Various devices are available for the measurement of variation and are usually included in every well-provided kit of statistical tools. Although these devices differ in detail, all fall into one of three broad categories corresponding to the *procedure on which they rest:* (1) *measurements of range* that include all or a specific percentage of the items; (2) measurements based on deviations of variates from a selected central value; and (3) measurements of heterogeneity in qualitative variables.

Some preliminary visual notion of dispersion in quantitative data may be obtained from graphic sketches, such as those presented in Figure 6.1.1. Thus, from the frequency graph we may discern whether the values are *symmetrically* dispersed around the *mean,* or whether the values above

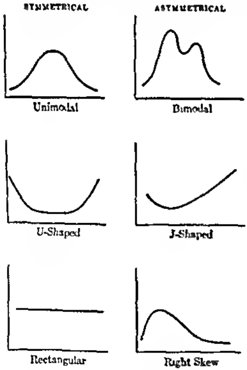


Figure 6.1.1 *Selected Patterns of Dispersion*

Table 6.1.1 *Computation of Selected Quantiles, Annual Family Income, U.S., 1956*

Total Money Income	Per Cent	Cumulated Per Cent
Under $500	3.2%	3.2%
$500– $999	3.3	6.5
$1,000– $1,499	4.4	10.9
$1,500– $1,999	4.5	15.4
$2,000– $2,499	5.1	20.5
$2,500– $2,999	5.1	25.6
$3,000– $3,499	6.2	31.8
$3,500– $3,999	6.3	38.1
$4,000– $4,499	8.0	46.1
$4,500– $4,999	6.9	53.0
$5,000– $5,999	13.7	66.7
$6,000– $6,999	9.8	76.5
$7,000– $9,999	15.6	92.1
$10,000–$14,999	5.9	98.0
$15,000–$24,999	1.5	99.5
$25,000 and over	0.5	100.0
Total	100.0%	
N =	43,445,000	

$$Md = 4{,}500 + \left(\frac{50 - 46.1}{6.9} \times 500\right)$$
$$= \$4{,}783$$

$$Q_1 = 2{,}500 + \left(\frac{25 - 20.5}{5.1} \times 500\right)$$
$$= \$2{,}941$$

$$Q_3 = 6{,}000 + \left(\frac{75 - 66.7}{9.8} \times 1{,}000\right)$$
$$= \$6{,}847$$

$$C_{10} = 1{,}000 + \left(\frac{10 - 6.5}{4.4} \times 500\right)$$
$$= \$1{,}398$$

$$C_{90} = 7{,}000 + \left(\frac{90 - 76.5}{15.6} \times 3{,}000\right)$$
$$= \$9{,}596$$

Source. U.S. Bureau of the Census, *Current Population Reports*, Series P–60, *Consumer Income*, No. 26, Table B, U.S. Government Printing Office, Washington, D.C., 1957.

which indicates that the middle 50 per cent of American family incomes are located on an interval slightly less than $4,000 in amount.

The interquartile range may be graphically displayed by plotting the quartiles on the base line of the graph (Figure 6.1.2), which reveals the extent to which the middle segment of the items are bunched around the median.

This graph also demonstrates that class intervals constructed so as to contain equal frequencies will generally not be identical in width. In this instance, the four intervals created by the quartiles vary enormously in width, although each of the four intervals contains exactly 25 per cent of the total frequency. Such a classification of items reverses the procedure of the orthodox frequency table, in which intervals are made equal and frequencies are left to vary. Here, we set the class frequencies equal and permit the class widths to vary. Both orderings are conveniently shown on the same graph.

A little reflection will indicate that there is no limit to the construction

Simple in conception and calculation, the range, like every other statistic, provides only limited information. Because its significance is relative to its location on the scale, it will generally be more serviceable when quoted in conjunction with its boundary points. Everyday usage recognizes the soundness of this principle in such expressions as "Prices on new cars will range from $4,000 to $5,000," or "Tomorrow's temperatures will range from a low of 42° to a high of 78°." A salary schedule that extends from $5,000 to $10,000 has the same absolute range as one that extends from $20,000 to $25,000, but they have very different connotations. In choosing a vacation site, it is not enough to know that the range in temperature is 30°; it is equally necessary to know the scale location of the extreme temperatures.

The range has the further characteristic that it disregards the pattern of variation between the extremes, and yet at times this pattern may be of greater import. The range of annual family incomes in the United States, which is greatly in excess of one million dollars, gives no clues to whether the incomes are compactly bunched in the middle, concentrated at one end, or uniformly spread over the entire scale.

Furthermore, in most observed distributions, the extreme values are infrequent, erratic, and unstable; hence, the over-all range, which rests exclusively on these extremes, may leave the impression of a greater volume of variation than actually exists. By basing the age range of college students on the singular 14-year-old prodigy and the 64-year-old mother who wishes to attend school with her grandchildren, we obtain a range of 51 years. But this result obscures the fact that most students differ from one another by only a few years, and the total range is therefore misleading as an index of variation.

Intermediate Ranges. This dependency on the almost unique extreme observations may be overcome by computing an intermediate range which excludes a minor fraction of the cases at either end, but which still includes a significant portion By basing it on the less exceptional items, the range acquires greater stability and dependability. One common practice is to take the difference between the 90th and 10th centiles, and thereby establish a range that includes the middle 80 per cent of the cases. Applying this procedure to the distribution of American family incomes in 1956 (Table 6.1 1) we obtain the 10–90 range:

$$C_{90} - C_{10} = \$9,596 - \$1,393 = \$8,193$$

A still more restricted range is the span between the first and third quartiles or the interval that subtends the middle 50 per cent of the items. It is naturally called the ***interquartile range.*** The interquartile range of family incomes is:

$$Q_3 - Q_1 = \$6,847 - \$2,941 = \$3,906$$

2. When are two or more ranges comparable?
3. How would you calculate the range when the ends of the frequency table are open?
4. Explain why the quartiles and the median practically never divide the total range into four equal intervals. In what type of curve will they partition the range equally?
5. The scores on a true–false sociology test of 100 questions ranged from 50 through 90. How would the range be affected if:
   (a) a bonus of 10 had been given to each student?
   (b) the poorest students had been successfully tutored?
   (c) an easier examination had been used? a more difficult exam?
   (d) 200 items had been used instead of 100?
6. Calculate the 9 decile points of the percentage distribution of family incomes (Table 6.1.1).

SECTION TWO

Variation as Measured by Arithmetic Deviations

It has been shown that the range, or any segment of it, will supply an impression of the span of a distribution. Although such a measure has some utility, especially when associated with its location, it represents merely the limits of variation, rather than the aggregate variation within those limits. It does not reflect the variation of the individual items, but merely the variation of the observed extremes, ignoring the many intermediate values. It may therefore be reiterated that the range measures the boundaries of scatter, but not the total amount of variation of the collectivity. Hence, some measure must be devised which will reflect the extent of diversity among all the items.

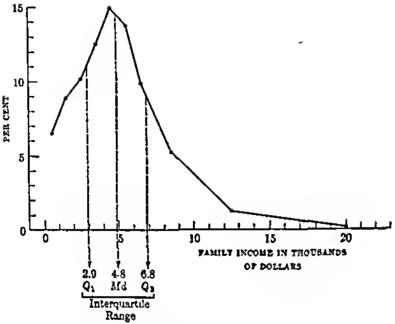


FIGURE 6.1.2 *Frequency Polygon, Annual Family Income, U.S., 1956*

of intermediate ranges, such as the 10–90 or the interquartile range. Any intermediate range will bring into clearer focus the relative degree of concentration or scatter among the items, particularly when viewed against the total range. The fact that the range of American family incomes is considerably greater than one million dollars, but that the interquartile range is only $3,906, and the 10–90 range is but $8,193, is suggestive of the essentially high degree of homogeneity among family incomes in the United States. Frequently, the use of strategically placed *quantiles* (the general term for partition values) will be an altogether satisfactory approach to the analysis of variation, and more complicated methods need not be pursued

QUESTIONS AND PROBLEMS

1. Define the following concepts:

Variability
Variation
Total Range (Range)
Intermediate Range
Interquartile Range
10–90 Range
Dispersion
Scatter
Spread
Kurtosis

more importantly, the sum of the arithmetic deviations around the median is less than from any other point of origin. Expressed in another way, the median is that point around which the arithmetic "errors" are least. This may be termed the *principle of minimal deviation.* The organization of an array around the median will therefore give us the most economical and tightly knit arrangement of items. Hence, the median would appear to be the more logical origin whenever the measure of variation is based on the absolute, or arithmetic, deviations.

Construction of a Deviational Measure: Average Deviation (AD). The mere sum of the arithmetic deviations is useless as an index of variation, since it will vary with the number of items in the distribution. For example, it will generally be large when there are 1,000 items, small when there are only 10 items. To eliminate this adventitious factor, we divide by N and thereby measure the deviation per case. This result is termed the *mean deviation* or *average deviation (AD).*

$$AD_{Me} = \frac{\Sigma|\text{d}|}{N} \tag{6.2.1}$$

where $|\text{d}|$ = deviation from median, sign ignored. Applying this formula to Table 6.2.1, we get:

$$\begin{aligned} AD_{Me} &= \frac{22}{5} \\ &= 4.4 \end{aligned}$$

The average deviation may also be based on the mean, in which case the above formula would read:

$$AD_{\bar{X}} = \frac{\Sigma|x|}{N} \tag{6.2.2}$$

where $|x|$ = a deviation from the mean, sign ignored. Thus,

$$\begin{aligned} AD_{\bar{X}} &= \frac{24}{5} \\ &= 4.8 \end{aligned}$$

which is larger than that based on the median. Hence, whenever skew is present, the median always serves best as the point of origin to express the amount of variation in the distribution.

Coefficient of Relative Variation (CRV).[*] It is now clear that any deviation takes on significance only when compared to its own origin, or norm. A variation of $2 in relation to a base of $10 conveys a different meaning

[*] Sometimes written *CV.* Broadly defined, a *coefficient* is a measure of relationship between two variables, expressed as a ratio, proportion, or percentage

It is also quite conceivable to think of death rates, student's grades, or teacher salaries in relation to a maximum or minimum — especially when they are in close proximity to these observed extremes. A teacher will perceive his salary as near the ceiling of his professional category. Nevertheless, it is more usual to employ a central value as a point of reference, because most values are actually in proximity to that central point. In other words, we tend to organize our observations around an average as a norm, on the assumption that this average is a representative value and therefore worthy of being used as a base of comparison. The only statistical issues still remaining are: (1) from what average to compute the deviations, and (2) how to summarize these deviations in a compact index.

The Choice of a Norm. Obviously, with symmetrical distributions that are unimodal, it makes little difference which average is used, since mode, median, and mean are equal. But concrete data often fall short of even approximate symmetry around a single peak. Hence, the problem is one of selecting a representative value in those distributions where the mode, median, and mean diverge one from another.

Unquestionably, many an observer informally employs the mode — the most frequent of his observations — as a base of comparison. But on theoretical grounds both the mean and the median can present somewhat stronger claims to the title of representativeness, and for that reason they are extensively employed in statistical calculations of this type.

As we have learned, the mean is the value from which the deviational variation on the two sides are in balance (Table 6.2.1); hence, the mean would appear to be the most reasonable point of origin of the deviations. However, the median's claim is at least equally impressive. The median is the position which divides the array into two equal parts; hence, and

Table 6.2.1 *Algebraic and Arithmetic Deviations from Mean and Median, Where $\bar{X} = 12$ and $Md = 10$*

VALUE	ALGEBRAIC DEVIATION		ARITHMETIC DEVIATION	
X	$X - \bar{X}$	$X - Md$	$\lvert X - \bar{X} \rvert$	$\lvert X - Md \rvert$
6	−6	−4	6	4
8	−4	−2	4	2
10	−2	0	2	0
15	3	5	3	5
21	9	11	9	11
60	0	10	24	22

Source Hypothetical

$$CRV = \frac{.37}{.92} \times 100 = 40\%$$

$$CRV = \frac{3.04}{8.76} \times 100 = 36\%$$

The results demonstrate that the two groups of states are approximately equal in relative variation; in fact, the South Atlantic group now appears to be even slightly more homogeneous than the New England states. This conclusion is independent of N and of scale location, both of which were *neutralized by this operation.*

The coefficient of relative variation is simple to compute, and particularly useful in comparative work, since it has the effect of norming for differences in absolute magnitudes and in substantive units of measure. It makes comparable sets of small and large values of the same kind, as well as values *that are qualitatively different.*

The *CRV* is not applicable, however, unless (1) the observed measures have a true zero, and (2) the scale intervals are equal — in short, unless we have ratio scales. It is accordingly not to be used to gauge the relative variation, for example, in measures of social distance, intelligence, and attitudes. The reasoning is as follows. The *CRV* is designed to standardize for differences in location of absolute measures, or more specifically, for differences in absolute magnitudes of central values such as the mean or median. But when an absolute zero is nonexistent, we must necessarily assign arbitrary values to a series of observations. Thus, a social-distance measure of 30 may represent the same objective fact as a measure of 120, according to the scale system arbitrarily set up by the investigator. Such scales are not anchored to an empirical zero which represents the absence of that phenomenon. If, then, locational measures are arbitrary, any standardization or correction would be devoid of meaning.

Questions and Problems

1. Define the following concepts:
   Arithmetic Deviation
   Algebraic Deviation
   Average Deviation (*AD*)
   Principle of Minimal Deviation
   Coefficient of Relative Variation (*CRV*)
   Absolute Zero
   Arbitrary Zero

2. When is it more logical to base the *AD* on the median rather than the mean?

3. Eleven houses are located on a street that runs due east and west. Which house is connected by the shortest aggregate distance to all others? Demonstrate graphically. How would the question be rephrased if there were ten houses?

from what it does in relation to a base of \$100. A variation of an inch in the length of the arm is no more than might be expected, but a variation of an inch in the length of the nose would be catastrophic. This principle is recognized and implemented in the *coefficient of variation*, which expresses the measure of variation as a percentage of its origin, be it mean or median. It thereby eliminates the extraneous factor of scale location. In the case of an AD based on the median, the formula would read:

$$CRV = \frac{AD_{Md}}{Md} \times 100 \tag{6.2.3}$$

If the AD is measured from the mean, we of course replace the median by the mean in Formula 6.2.3.

Let us employ this procedure to compare the relative variation between homicide rates in the New England and the South Atlantic states (Table 6.2.2). After comparing the raw AD's of New England (.37) and the South Atlantic states (3.04), we might conclude that the New England states are much more homogeneous, since their average divergence from the norm is so much narrower. But such a conclusion would be premature, for it is possible that, *relative to their origin, the deviations within the South Atlantic* group are no larger than those within the New England group. To determine whether this is the case, we express each AD as a percentage of its base median:

Table 6.2.2 *Computation of Average Deviation, Homicide Rates per 100,000 Population, New England and South Atlantic States, 1952*

NEW ENGLAND	RATE	$\lvert d \rvert$	SOUTH ATLANTIC	RATE	$\lvert d \rvert$
Maine . .	.82	.10	Delaware	3.29	5.47
Massachusetts .	.83	.09	West Virginia....	6.01	2.75
Vermont . .	.85	.07	South Carolina...	7.59	1.17
Rhode Island	.98	.06	Maryland........	8.56	.20
Connecticut .	1.78	.86	Virginia..........	8.95	.19
New Hampshire	1.95	1.03	Florida..	9.97	1.21
		$\Sigma = 2.21$	North Carolina ..	11,19	2.43
			Georgia...... ..	20.67	11.91
					$\Sigma = 25.33$
$Md = .92$ $AD = .37$			$Md = 8.76$ $AD = 3.17$		

Source U S Department of Justice, *Uniform Crime Reports, Annual Bulletin, 1952, U.S.* Government Printing Office, Washington, D C., 1952.

tions from the median is minimal, so the sum of the squared deviations from the mean is also minimal. This is an exemplification of the *principle of least squares*, which is one of the most venerable and vital principles in all statistics, known and practiced for 150 years.

The technique of squaring deviations may, at first glance, seem unnecessarily circuitous and superfluous. If variation can be satisfactorily measured by simple deviations, what additional information and insight can be gained by squaring them? A completely adequate answer to this question will be possible only at a later stage of the student's statistical studies. It must here suffice to state that the practical utility of such a measure is incomparably greater than that of the *AD*, which is less frequently called into use. The squared deviations may be expressed in several ways, each of which serves its own purpose: *sum of squares* (*SS*), *variance* (*V*), and *standard deviation* (*SD*).

The Sum of Squares (*SS*). Just as we computed the sum of the arithmetic deviations from the median and the mean in Table 6.2.1, we may now compute the sum of the *squared* deviations from both averages (Table 6.3.1). Although the sum of simple arithmetic deviations is minimal from the median, we see in Table 6.3.1 that the sum of the *squares* of deviations is minimal when these deviations are measured from the mean.

This result may seem paradoxical. Mathematical proof for this apparent inconsistency cannot be offered here, but a partial explanation lies in the fact that squaring is a geometric computation which has the effect of weighting disproportionately the deviations as they increase in magnitude. Since the mean equalizes the negative and positive deviations, it operates to reduce to the limit the relative frequency of the large deviations, and thereby minimizes the *SS*, usually written Σx^2.

Table 6.3.1 *Sum of Squares of Deviations from Mean and Median, where* $\bar{X} = 12$, $Md = 10$

X	$\lvert X - \bar{X} \rvert$	$\lvert X - Md \rvert$	$(X - \bar{X})^2$	$(X - Md)^2$
6	6	4	36	16
8	4	2	16	4
10	2	0	4	0
15	3	5	9	25
21	9	11	81	121
60	24	22	$SS = 146$	$SS = 166$

Source. Hypothetical

4. Compute the *AD* from the median for the following set ($N = 20$):

7	4	6	5	10
9	13	6	6	7
4	21	11	0	10
16	8	4	12	6

   (a) How would the *AD* be affected if each value were increased by 10?
   (b) How would the *CRV* be affected?

5. Table 6.2.3 shows the percentage of each repertoire (e.g., Boston Orchestra, Chicago Orchestra, etc.) devoted to each of six leading composers. Calculate the median percentage for each of the composers, the average deviation based on the median, and the *CRV* for each composer. Rank the composers according to their *CRV*, and interpret.

Table 6.2.3 *Composer Representation, Orchestral Repertoires, 1940–1944*

COMPOSER	ORCHESTRA							
	Boston	*Chicago*	*Cincinnati*	*Cleveland*	*Minneapolis*	*New York Philharmonic*	*Philadelphia*	*St. Louis*
Bach	9.8	4.9	2.0	1.8	2.3	2.9	3.3	1.4
Beethoven .	9.6	10.9	8.1	11.0	11.4	10.3	10.7	9.3
Berlioz . . .	2.5	1.6	1.1	0.7	2.6	2.3	0.6	0.1
Brahms . .	9.2	9.4	7.4	11.2	11.4	9.8	10.8	11.9
Prokofiev . .	1.2	1.2	0.5	1.1	1.1	1.1	1.8	1.0
Tschaikowsky	4.5	5.9	7.3	9.6	2.8	5.4	6.5	9.3

Source John H Mueller, "The Measurement of Aesthetic Folkways," *American Journal of Sociology*, LI 1946, p 278 (adapted with original data from author).

SECTION THREE

Variation as Measured by Squared Deviations

Principle of Least Squares. The measurement of variation by simple arithmetic deviations from a central value is a straightforward procedure. If a simple statement of dispersion were all that was wanted, the easily comprehended *AD* would serve quite well. However, variation may be, and is even more commonly, measured by squared deviations, which are universally taken from the mean of the series. The fundamental logic of this procedure is identical with that which argues for the simple deviations from the median: the principle of best fit. Just as the sum of the deviations from the median is minimal, so the sum of the squared deviations from the mean is also minimal. This is an exemplification of the *principle of least squares*, which is one of the most venerable and vital principles in all statistics, known and practiced for 150 years.

The technique of squaring deviations may, at first glance, seem unnecessarily circuitous and superfluous. If variation can be satisfactorily measured by simple deviations, what additional information and insight can be gained by *squaring them? A completely adequate answer to this* question will be possible only at a later stage of the student's statistical studies. It must here suffice to state that the practical utility of such a measure is incomparably greater than that of the *AD*, which is less frequently called into use. The squared deviations may be expressed in several ways, each of which serves its own purpose: *sum of squares* (*SS*), *variance* (*V*), and *standard deviation* (*SD*).

The Sum of Squares (*SS*). Just as we computed the sum of the arithmetic deviations from the median and the mean in Table 6.2.1, we may now compute the sum of the *squared* deviations from both averages (Table 6.3.1). Although the sum of simple arithmetic deviations is minimal from the median, we see in Table 6.3.1 that the sum of the *squares* of deviations is minimal when these deviations are measured from the mean.

This result may seem paradoxical. Mathematical proof for this apparent inconsistency cannot be offered here, but a partial explanation lies in the fact that squaring is a geometric computation which has the effect of weighting disproportionately the deviations as they increase in magnitude. Since the mean equalizes the negative and positive deviations, it operates to reduce to the limit the relative frequency of the large deviations, and thereby minimizes the *SS*, usually written Σx^2.

Table 6.3.1 *Sum of Squares of Deviations from Mean and Median, where $\bar{X} = 12$, $Md = 10$*

X	$\lvert X - \bar{X} \rvert$	$\lvert X - Md \rvert$	$(X - \bar{X})^2$	$(X - Md)^2$
6	6	4	36	16
8	4	2	16	4
10	2	0	4	0
15	3	5	9	25
21	9	11	81	121
60	24	22	$SS = 146$	$SS = 166$

Source: Hypothetical

Variance (V). Although the sum of squares is necessarily employed in various statistical procedures, it is not a meaningful index of variation — and for the same reason that the sum of the arithmetic deviations around any average is not. A meaningful index of variation is obtained by taking the mean of the sum of squares, and thereby eliminating the variable *factor of frequency. This result, called the variance, and most commonly* symbolized σ^2, corresponds procedurally to the average deviation. The formula is:

$$\sigma^2 = \frac{\Sigma x^2}{N} \tag{6.3.1}$$

The Standard Deviation (SD), or Sigma (σ). Since the variance is derived from the squared deviations, it is not a linear measure. If such a measure is required, as it usually is, we have only to unsquare the variance in order to restore the linear state. In this form, it is known as the *standard deviation*, or *sigma*, symbolized by *SD*, or the lower-case Greek σ.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \tag{6.3.2}$$

As such, it is used as a measure of variation quite analogous to the *AD*, from which it differs primarily in the fact that the deviations were squared, and their mean unsquared. However, the unsquaring does not cancel the total effect of previous squaring; the weighting effect in part remains. If it had been completely nullified, the result would have been the simple *AD* from the mean.

Table 6.3.2a *Computation of Variance and Standard Deviation, Homicide Rates, South Atlantic States, 1952*

STATE	X	x	x^2
Delaware........ . . .	3	−6.5	42.25
Florida	10	0.5	.25
Georgia...... . . .	21	11.5	132.25
Maryland........ ..	8	−1.5	2 25
North Carolina.. . .	11	1.5	2.25
South Carolina	8	−1.5	2.25
Virginia..	9	−0 5	.25
West Virginia.	6	−3 5	12.25
	$\Sigma = 76$	0.0	$SS = 194.00$
	$\bar{X} = 9.5$		

$$V = \sigma^2 = \frac{\Sigma x^2}{N} = \frac{194}{8} = 24.25$$

$$SD = \sigma = \sqrt{\tfrac{194}{8}} = \sqrt{24.25} = 4.9$$

Computation of the Standard Deviation. (*a*) *Ungrouped Data.* In principle, the standard deviation is slightly more complicated than is the average deviation, involving only the additional steps of squaring the deviations and unsquaring their mean. This is illustrated in Table 6.3.2a.

The calculation of the standard deviation on a large aggregate would be extremely laborious if each deviation had to be individually measured and squared. There are, however, computing formulas which greatly simplify such calculations. These formulas all involve one or more coding operations, similar to those employed in the case of the mean, although these operations are masked from view by the terms of the formula. A particularly expeditious method for finding the standard deviation of ungrouped data performs all calculations on the raw observed values themselves.

In the table below are presented the computing formulas that comprise this method alongside the basic formulas in order to emphasize the correspondence between them. To illustrate the computing routines, they are applied to the homicide rates of the same eight South Atlantic states (Table 6.3.2b).

Name	*Notation*	*Basic Formula*	*Computing Formula*	
Sum of Squares	$N\sigma^2$	Σx^2	$\Sigma X^2 - \frac{(\Sigma X)^2}{N}$	(6.3.3)
Variance	σ^2	$\frac{\Sigma x^2}{N}$	$\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2$	(6.3.4)
Standard Deviation	σ	$\sqrt{\frac{\Sigma x^2}{N}}$	$\sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$	(6.3.5)

Table 6.3.2b *Computation of Variance and Standard Deviation by Coding, Homicide Rates, South Atlantic States, 1952*

X	X^2
3	9
10	100
21	441
8	64
11	121
8	64
9	81
6	36
76	916

$$\Sigma x^2 = 916 - \frac{(76)^2}{8} = 194$$

$$\sigma^2 = \frac{\Sigma x^2}{N} = \frac{916}{8} - \left(\frac{76}{8}\right)^2 = 24.25$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{916}{8} - \left(\frac{76}{8}\right)^2}$$

$$= \sqrt{\frac{194}{8}} = 4.9$$

These yield exactly the same answer as that obtained by the basic formula. While the calculation of the *SD* will thus normally be executed on the basis of the computing routines, its interpretation will always be in terms of the basic definitional formula

(*b*) *Grouped Data.* To obtain the *SD* of grouped data, simple logic would suggest that we measure the deviations of the class midpoints from the mean and weight these according to the frequencies in the respective intervals:

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}} \tag{6.3.6}$$

where f = class frequency
x = deviation of class midpoint from mean

But these cumbersome arithmetic operations do not match the seductive simplicity of the logic involved. Hence, we once again resort to coding midpoints by interval deviations from the guessed mean, as was done in the computation of the mean of grouped data (see Chapter 5, pp. 136–138). The augmented worksheet requires only one additional column in order to provide for the squared deviations (Column 5, Table 6.3.3). In the table below, we have once again placed computing formulas alongside the basic formulas to emphasize their equivalence, and in Table 6.3 3 we apply the computing formula of the *SD* to the tabulation of suicide rates.

Name	*Notation*	*Basic Formula*	*Computing Formula*	
Sum of Squares	$N\sigma^2$	$\Sigma f x^2$	$i^2\left\{\Sigma f x'^2 - \frac{(\Sigma f x')^2}{N}\right\}$	(6.3.7)
Variance	σ^2	$\frac{\Sigma f x^2}{N}$	$i^2\left\{\frac{\Sigma f x'^2}{N} - \left(\frac{\Sigma f x'}{N}\right)^2\right\}$	(6.3 8)
Standard Deviation	σ	$\sqrt{\frac{\Sigma f x^2}{N}}$	$i\sqrt{\frac{\Sigma f x'^2}{N} - \left(\frac{\Sigma f x'}{N}\right)^2}$	(6.3 9)

The Coefficient of Relative Variation (*CRV*). Like the *AD*, the standard deviation may be converted into a measure of relative variation by *norming it* on its own origin, namely the mean.

$$CRV = \frac{\sigma}{\bar{X}} \times 100 \tag{6.3.10}$$

Applying it illustratively to the suicide data, where $\bar{X} = 12.5$ and $\sigma = 5.4$, we obtain:

$$\begin{aligned} CRV &= \frac{5.4}{12.5} \times 100 \\ &= 43\% \end{aligned}$$

Table 6.3.3 *Computation of Standard Deviation by Coding, Grouped Data, Suicide Rates, 107 Large U.S. Cities*

(1)	(2)	(3)	(4)	(5)	
X	f	x'	fx'	fx'^2	
3– 5	6	−3	−18	54	$\sigma = i\sqrt{\frac{\Sigma fx'^2}{N} - \left(\frac{\Sigma fx'}{N}\right)^2}$
6– 8	18	−2	−36	72	
9–11	29	−1	−29	29	$= 3\sqrt{\frac{346}{107} - \left(\frac{-18}{107}\right)^2}$
12–14	24	0	0	0	
15–17	13	1	13	13	$= 3\sqrt{3.23 - .03}$
18–20	7	2	14	28	$= 3\sqrt{3.20}$
21–23	4	3	12	36	$= 3(1.79)$
24–26	4	4	16	64	$= 5.37$
27–29	2	5	10	50	
	107		−18	346	

As previously stated, such a measure is not likely to be used in isolation from other comparable measures, except occasionally to indicate the extent of scatter in relation to the magnitude of the mean. Thus, a small sigma value relative to a large mean represents great homogeneity of the data, and consequent typicality of the mean, which may be an important piece of information under certain circumstances. A mean of $125, with a σ of $5, giving a *CRV* of 4 per cent, is more representative of the array than a mean of $125 with a σ of $25 and a *CRV* of 20 per cent. A *CRV* of zero indicates no variation at all, and that the mean is utterly typical.

However, the *CRV* is more likely to be used in outright comparisons between series of related data. The relative variation in wages in the East, for example, may be less than in California. In the measure of public taste for a given composer, a high *CRV* would result from great disagreement; a low coefficient on the contrary would reflect a tendency toward consensus in taste.

It is phenomena like these that the *CRV* is designed to measure in the form of a simple, precise, and compact index. It should be stated here that, to the extent that the *SD* exaggerates the variation around the mean, the *CRV* correspondingly exaggerates the relative variation. But whether the *CRV* is based on the *SD* or *AD*, it is of course inapplicable when the observed measures have no absolute zero.

Characteristics of the Standard Deviation. For almost a century, the *SD* has been one of the most powerful tools in statistical analysis, not only as a measure of dispersion, but also as an ingredient in more complicated computations. There are two interlocking characteristics of the sigma which bear the credit for the efficacy of this tool. First, the standard deviation reflects the magnitude of every variate of the series twice over: (a) the point of origin from which the deviations are measured ($\bar{X}$) is itself an arithmetic reflection of the magnitudes of all the variates, and (b) every magnitude, as magnitude, is represented arithmetically in the squared deviations. Secondly, the squaring of the deviations constitutes a mathematically *legitimate method of clearing the signs*. While the *AD*, measured from the median, "ignores" the signs, the standard deviation "clears" the signs mathematically; consequently, there is no breaking of the chain of mathematical operations.

However, for descriptive and terminal purposes, the expedient of ignoring the signs in the calculation of the *AD* is quite legitimate. Because it measures the deviations directly, without squaring, it is smaller than the *SD*, which inflates the deviations disproportionately by squaring. However, as already implied, ignoring the signs disqualifies the *AD*, irrespective of its origin, from participating in further algebraic computations.

So "standard" has the standard deviation become that, by force of habit, it is employed as a simple, literal description of variation, even when the *AD* would be more faithful to the absolute deviations. This would be the case whether the deviations were measured from the median or mean, both of which are sanctioned by general practice.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Principle of Least Squares
   Sum of Squares
   Variance
   Standard Deviation
   Basic Formula
   Computing Formula
2. Explain why the standard deviation is ordinarily used as a measure of dispersion instead of the variance.
3. Why is the *SD* larger than the *AD*?
4. Using small sets of illustrative data, verify that the sum of the squares is least from the mean.
5. Is it possible for an array to have more than one *SD*?
6. Is it possible to calculate the *SD* of qualitative data?
7. From an examination of the basic formula for the *SD*, explain why "unsquaring" does not completely nullify the effect of the previous squaring.
8. Compute the *SD of the series given in Problem 4*, p. 162. Add a constant (e.g., 20) to each value and compute the *SD* of this transformed series. How is the *SD* affected?

9. If the mean age of a group of college students is 20 years, with a standard deviation of 2, what will be the mean and *SD* of that group 20 years later? The *CRV*?

10. Compute the *SD* of the frequency distribution of delinquency rates (Problem 11, p. 47) by the computing formula for grouped data.

11. Discuss the aptness of (a) the range, (b) the interquartile range, (c) the average deviation, and (d) the standard deviation for purposes of describing the dispersion of the following distribution (Figure 6.3.1):

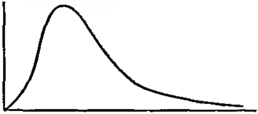


Figure 6.3.1 *Skewed Distribution*

Section Four

The Normal Distribution as a Pattern of Variation

The Concept of the Normal Frequency Distribution. One of the most important applications of the *SD* lies in its description of the normal curve. *In fact, it may be claimed that as a measure of dispersion, the standard deviation has meaning only insofar as the pattern of variation is normal.* Hence, an adequate discussion of the standard deviation must specify its fundamental relation to the normal curve, the most illustrious of all patterns of statistical variation.

The history of the normal curve dates back to 1733, when Abraham de Moivre first established it in the course of his investigation into games of chance. Later, in his more pious moments, he contended that it was a manifestation of divine order in the universe.

Since that time, it has served various purposes. To astronomers, it described the distribution of measurement errors around the "true" value; hence, it was often characterized as *the curve of error*. The Belgian statistician Quételet, during the 1830's, was the first to apply this curve to social, psychological, and anthropometric data. His own concept, *l'homme moyen* — the average man — which for him described the norm, flanked by "nature's errors," is the logical basis for the now prevalent terminology of "the normal distribution." The curve was also employed to represent a sampling distribution of all possible sample values, which is today one of its most significant uses — *a topic which will be treated in a later chapter.*

Quételet was of the opinion that social and moral data tend to array themselves on the normal curve, which is thereby given the sanction of nature. However, the "normal curve" no longer carries this eulogistic connotation. Today it would be a mistake to consider non-normal distributions as unusual or unnatural. The distribution of raw, empirical data may naturally conform to any one of a number of curves. Nevertheless, as a statistical model in fitting nature's variation, the normal curve is still unrivaled in scope of application, notwithstanding the fact that it does not possess the universality which Quételet attributed to it.

Characteristics of the Normal Curve. By definition, the ideal normal distribution contains an infinite number of cases, is unimodal and symmetrical, and unbounded at either end. Consequently, mean, median, and mode are identical in value, and divide the array into two equal parts. The graphic version of this distribution is a smooth, bell-shaped curve, with a characteristic slope that never touches the base line. Starting at its peak, the curve falls more and more rapidly up to the point of inflection, and then gradually levels off, extending indefinitely in either direction. This point of inflection is exactly one standard deviation distant from the mean origin. Graphically, then, the standard deviation is the linear distance along the base line from the mean to the ordinate defining the point of in-

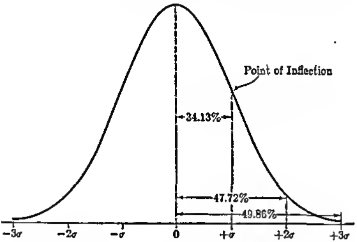


FIGURE 6.4.1 *Normal Curve, Standard Deviation, and Normal Areas*

flection. This unit has been almost universally employed in measuring distances from the mean since it was first approvingly launched by Karl Pearson.

For every possible sigma distance from the mean, there is necessarily a corresponding percentage of area, or frequency; hence, sigma points serve as convenient measures of location. If we travel one sigma from the mean, we leave behind 34 per cent of the items. But if we continue on to the second sigma point, we would not have put behind 68 per cent of the items, but rather only about 48 per cent, owing to the steady decline in density as the distance from the mean becomes greater. Since the distribution is symmetrical, approximately two-thirds (68.26 per cent) of the cases are included within the interval that extends from 1 *SD* below to 1 *SD* above the mean; 95 per cent (95.44 per cent) fall within 2 *SD*'s on either side; and practically 100 per cent (99.72 per cent) fall within 3 *SD*'s on either side of the mean.

Similarly, it is possible to determine the proportion of cases between the mean and any other sigma value on the base line. Because this type of information has so many applications, reference tables have been prepared for the convenience of statistical workers. The most common of these tables, the *Table of Normal Areas*, gives the proportion of the total area or items under the curve between the mean and selected multiples of the standard deviation. A systematic condensation of that table is reproduced below. Like the more complete source, from which it has been taken, it

Table 6.4.1 *Table of Normal Areas*

x/σ	AREA	x/σ	AREA	x/σ	AREA
.1	.0398	1.1	.3643	2.1	.4821
.2	.0793	1.2	.3849	2.2	.4861
.3	.1179	1.3	.4032	2.3	.4893
.4	.1554	1.4	.4192	2.4	.4918
.5	.1915	1.5	.4332	2.5	.4938
.6	.2257	1.6	.4452	2.6	.4953
.7	.2580	1.7	.4554	2.7	.4965
.8	.2881	1.8	.4641	2.8	.4974
.9	.3159	1.9	.4713	2.9	.4981
1.0	.3413	2.0	.4772	3.0	.4986

Source: See Table I, Appendix, p. 418.

describes only one-half of the distribution — which suffices, of course, since the distribution is symmetrical.

Examination of the table indicates, for example, that approximately

.4953 of the items lie within 2.6σ of the mean, so that practically 99 per cent lie within 2.6σ on either side of the mean. Consequently, for all practical purposes, the normal distribution has a range of six sigmas.

The Standard Deviate. The foregoing description of the normal curve is an abstract one, given in terms of the standard deviation measured from the mean as an origin. As such, the unit of measure is independent not only of diverse measurement systems, but also of the concrete values themselves. It makes no difference whether we are dealing with incomes of a hundred *or* a *million* dollars, with *durations of ten seconds or ten years*, or with varying intensity of attitude.

But sets of data always come to us as raw measures and give every appearance of being non-comparable. We cannot readily compare, for example, teachers' salaries and years of service, even though both variables may be normal in their distributions. The solution to this problem of comparability, of course, lies in converting the raw measures into sigma units of measure, which *are* comparable. We express raw deviations from the respective means as multiples of their standard deviations. Hence, such measures are called *standard deviates.*[*] They are conventionally symbolized *z*.

$$z = \frac{X - \bar{X}}{\sigma} \qquad (6.4.1)$$

$$= \frac{x}{\sigma}$$

By plotting absolute and sigma scales on the base line of a normal frequency graph, it is possible to display visually the equivalence of raw measures and standard deviates. It becomes plain, for example, that a teacher's salary of $6,000 coincides with a standard deviate of +1.00; both values represent the same objective fact. Moreover, by this device of multiple scales, it is possible to exhibit conveniently the essential identity between normally distributed variables which are seemingly very different. Thus, the illustrative graph makes clear that a *teacher's salary of* $6,000 is statistically identical with 16 years of teaching service, both lying one sigma from the mean.

This transformation to standard form, which at first may seem awkward, will gradually become second nature to every student of statistics, since it finds such a wide variety of application. Every student, whether or not he is familiar with statistical processes, well knows that raw total scores of, say 300 and 500 in English and Mathematics, may or may not be equivalent in relative grade value. But upon discovering both to be two sigmas above the mean, the statistically trained student will correctly judge them

[*] Synonyms include: *standard measure*, *standard score*, *z-measure*, *relative deviate*, and *sigma value*.

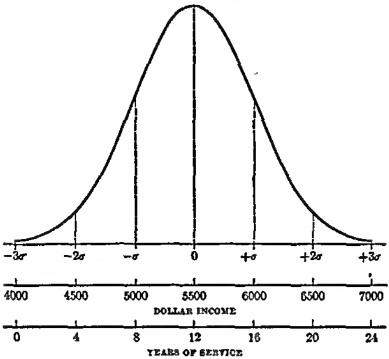


Figure 6.4.2 *Teacher Income and Years of Service as Sigma Units (Hypothetical Data)*

to be identical, since in both instances 98 per cent of the grades are presumed to be lower. Thus, the *z*-measure serves to establish the relative position of an item in a given array, and thereby renders corresponding items in two or more normal arrays comparable.

Calculation of the Standard Deviate. To convert any series of values into standard deviates, it is first necessary to compute the mean and the *SD* of the series. Such computations are carried out on the data shown in Table 6.4.2, which have purposely been curtailed for ease of comprehension. The first variate is 1, which deviates by -3 from the mean value of 4. Since the $SD = 2$, this variate deviates by -1.5σ. That is, the value 1 translated into standard form is equal to -1.5. The remaining standard measures are similarly calculated.

These *z*-measures enable us to fix the relative position of each item in the array by referring to the table of normal areas, or frequencies. Thus, from that table we read that 1.5σ, measured from the mean, corresponds to approximately 43 per cent of the total frequency, which leaves about 7 per cent of the items above that point. Although this latter figure cannot

Table 6.4.2

Computation of Standard Deviates

X	x	$\frac{x}{\sigma} = z$
1	-3	-1.5
3	-1	$-.5$
4	0	0
5	1	.5
7	3	1.5
20	0	
	$\bar{X} = 4$ $\sigma = 2$	

be read directly from the table, it may readily be calculated by subtracting the table entry from 50 per cent, as in the above example. It should be emphasized that, although sigmas may be mechanically computed for any distribution, whether normal or not, the foregoing interpretation of sigma measures is valid only for the normal distribution.

Choice of Measure of Variation. The z-measure is not the only standard measure in the most general sense of that term. Centiles, such as the median and the quartiles, are also independent of the concrete data and the measuring scales, and are therefore equally deserving of the title "standard measure" In fact, since they are more generally applicable, they are perhaps even more deserving of that appellation, because their meaning does not depend on the form of the distribution, as is true of z-measures. This is merely one of many instances of a generic term being taken over in a specific setting and assigned a specialized meaning. We therefore remind the student that each type of measure of variation has its appropriate function, and that familiarity with both its statistical characteristics and the substantive problem is necessary for an informed choice.

When the data are quantitative, the choice may depend on one or more of the following overlapping criteria: (1) the objective of the inquiry; (2) whether the measure is used as a terminal description or as a preliminary to further computations; (3) the pattern of the distribution, whether symmetrical or skewed; (4) the degree of completeness of the source data.

If our limited objective is only a quick impression of the scatter or density, the range may be quite satisfactory. However, a more reliable impression of the essential compactness of a distribution is conveyed by the intermediate ranges, such as the interquartile range, or the 10–90 decile range. These measures have the added versatility that their interpretation is completely independent of any particular pattern of distribution. In this sense, they are "distribution-free"

All summary deviational measures, of course, express the amount of deviation from a central value. Therefore, they are useful companion pieces of information to the statement of the central value itself. If a plain and simple measure of the volume of scatter around a central value is desired, the average deviation recommends itself as the most forthright. The standard deviation may be used for the same purpose, but it will not serve as well: in the first place, it is less comprehensible because of its more complicated derivation; and, secondly, the distortion produced by squaring reduces its descriptive value. In spite of this, there are those who adhere more or less indiscriminately to the standard deviation. This convention seems to derive from the general prestige of the sigma, and the habit of its employment for other purposes, rather than from a scrupulous examination of its adequacy as a descriptive device.

However, the *SD*, and secondarily the *AD*, carry a supplementary function of locating the relative position of an item within the distribution with the aid of the table of normal areas. Obviously in such instances a normal distribution must be assumed. If the pattern of distribution is not at least approximately normal, one should consider the alternative employment of centiles and other positional measures in the place of the sigma. When, however, more advanced computations are anticipated, the sigma is inescapable. It is used as a standard unit of measure in comparing otherwise non-comparable data, and it is presupposed in correlation and measures of sampling reliability.

To the extent that empirical sociological data are not normally distributed, the sigma may possibly have less utility in sociology than in the descriptive statistics of other disciplines; but this "handicap" does not carry over into sampling, for reasons which will be made clear in subsequent chapters.

Questions and Problems

1. Define the following concepts:
   Normal Frequency Distribution
   Normal Area
   Normal Ordinate
   Standard Deviate
   z-Measure
   Bell-Shaped Distribution

2. Explain in what sense the "normal" curve is normal.

3. A given variate lies one *SD* above its mean. Express this variate as a standard deviate.

4. The range of a group of adult weights is 100 pounds; the range of their heights is 12 inches. How might these ranges be made comparable? What data would be required?

Table 6.4.2

Computation of Standard Deviates

X	x	$\frac{x}{\sigma} = z$
1	-3	-1.5
3	-1	$-.5$
4	0	0
5	1	.5
7	3	1.5
20	0	
$\bar{X} = 4$ $\sigma = 2$		

be read directly from the table, it may readily be calculated by subtracting the table entry from 50 per cent, as in the above example. It should be emphasized that, although sigmas may be mechanically computed for any distribution, whether normal or not, the foregoing interpretation of sigma measures is valid only for the normal distribution.

Choice of Measure of Variation. The z-measure is not the only standard measure in the most general sense of that term. Centiles, such as the median and the quartiles, are also independent of the concrete data and the measuring scales, and are therefore equally deserving of the title "standard measure." In fact, since they are more generally applicable, they are perhaps even more deserving of that appellation, because their meaning does not depend on the form of the distribution, as is true of z-measures. This is merely one of many instances of a generic term being taken over in a specific setting and assigned a specialized meaning. We therefore remind the student that each type of measure of variation has its appropriate function, and that familiarity with both its statistical characteristics and the substantive problem is necessary for an informed choice.

When the data are quantitative, the choice may depend on one or more of the following overlapping criteria: (1) the objective of the inquiry; (2) whether the measure is used as a terminal description or as a preliminary to further computations; (3) the pattern of the distribution, whether symmetrical or skewed; (4) the degree of completeness of the source data.

If our limited objective is only a quick impression of the scatter or density, the range may be quite satisfactory. However, a more reliable impression of the essential compactness of a distribution is conveyed by the intermediate ranges, such as the interquartile range, or the 10–90 decile range These measures have the added versatility that their interpretation is completely independent of any particular pattern of distribution. In this sense, they are "distribution-free."

All summary deviational measures, of course, express the amount of *deviation from a central value.* *Therefore, they are useful companion* pieces of information to the statement of the central value itself. If a plain and simple measure of the volume of scatter around a central value is desired, the average deviation recommends itself as the most forthright. The standard deviation may be used for the same purpose, but it will not serve as well: in the first place, it is less comprehensible because of its more complicated derivation; and, secondly, the distortion produced by squaring reduces its descriptive value. In spite of this, there are those who adhere more or less indiscriminately to the standard deviation. This convention seems to derive from the general prestige of the sigma, and the habit of its employment for other purposes, rather than from a scrupulous examination of its adequacy as a descriptive device.

However, the *SD*, and secondarily the *AD*, carry a supplementary function of locating the relative position of an item within the distribution with the aid of the table of normal areas. Obviously in such instances a normal distribution must be assumed. If the pattern of distribution is not at least approximately normal, one should consider the alternative employment of centiles and other positional measures in the place of the sigma. When, however, more advanced computations are anticipated, the sigma is inescapable. It is used as a standard unit of measure in comparing otherwise non-comparable data, and it is presupposed in correlation and measures of sampling reliability.

To the extent that empirical sociological data are not normally distributed, the sigma may possibly have less utility in sociology than in the descriptive statistics of other disciplines; but this "handicap" does not carry over into sampling, for reasons which will be made clear in subsequent chapters.

Questions and Problems

1. Define the following concepts:
   Normal Frequency Distribution
   Normal Area
   Normal Ordinate
   Standard Deviate
   z-Measure
   Bell-Shaped Distribution

2. Explain in what sense the "normal" curve is normal.

3. A given variate lies one *SD* above its mean. Express this variate as a standard deviate.

4. The range of a group of adult weights is 100 pounds; the range of their heights is 12 inches. How might these ranges be made comparable? What data would be required?

Section Five

Variation of Qualitative Variables

Can Qualitative Variation Be Measured? The variation of qualitative variables cannot be measured in the same manner as that of quantitative data. Qualitative variables do not exist in magnitudes, and are not ranged on a continuum with a zero origin; there is no central value, or total or intermediate ranges. Hence, there are no deviations or, of course, average deviations.

But that is not to say that a group of qualitative items are necessarily identical, that there is only homogeneity and no heterogeneity. Two items differ when they do not possess the same attribute. Even though these differences are qualitative, it is still possible to devise some measure which effectively summarizes them. But instead of measuring magnitudes, we count differences.

Now it is a truism that the greater the number of differences among a set of items, the more heterogeneous is the aggregate, and consequently the more variation within it. Similarly, the smaller the number of differences, the greater the homogeneity within it, and the less the variation. There can be, for example, no sex difference at a stag party; but, in mixed company, there will always be a smaller or larger number of sex differences between individuals, depending on the sex ratio of that group. It appears reasonable, therefore, to base an index of qualitative variation on the total number of differences among the items in the given set. It is only a question of (1) how to compute the total number of qualitative differences, and (2) how to convert this total into a meaningful index.

To find the total number of differences, we count the differences between each item and every other item and sum these observed differences. For example, in a set of six boys and six girls, each of the six boys will differ in attribute from each of the six girls, thereby making a total of 36 sex differences. If there were nine boys and three girls, each of the nine boys would differ from each of the three girls, producing 27 differences. In a group of 12 boys, the obvious result of no differences would be obtained by multiplying 12 by zero.

Evidently, the procedure for determining the total number of actual differences reduces to the following rule: multiply every attribute frequency by every other attribute frequency and sum these products. For example, in a set of four Catholics, five Protestants, and six Jews, there would be: $(4 \times 5) + (4 \times 6) + (5 \times 6) = 74$ differences.

Index of Qualitative Variation (IQV). However, these observed differences take on meaning only in relation to the maximum possible number of

5. On the base line of a normal curve, sigma units are equal. Would this be the case for a skewed curve? Do quartile points divide the range of a normal distribution into equal intervals?

6. Plot a normal curve as follows:
   (a) From the zero origin placed at the midpoint of the horizontal axis, mark off three arbitrary sigma units in both directions.
   (b) Divide each sigma unit into tenths.
   (c) Above each such division point, plot the corresponding ordinate (Table II, Appendix, p. 419).
   (d) Draw a smooth curve through these ordinate points.
   (e) From the resultant figure, summarize what seem to be the main features of the normal curve.
   (f) Is it possible to determine visually from a graph, whether or not a distribution is normal?

7. Use Table I of the Appendix (p. 418) to determine the proportion of the normal area lying between the mean and following standard deviates:

   $\pm .67$ ± 1.96 ± 2.33 ± 3.00
   ± 1.00 ± 2.00 ± 2.58

8. Find the proportion of cases between each pair of sigma points on the base line of a normal curve. Represent results graphically.

   0.3 to 1.6 1.1 to 1.2 −2.58 to +2.58
   0.3 to −1.6 0 0 to 1.0 1.5 to 3 0
   .1 to .2 1.0 to 2 0 −2.3 to +2.3

9. Between what two sigma points on the base line of the normal curve do the middle 50 per cent of the cases lie?

10. Explain why the proportion of cases between zero and 1.0 sigma is not equal to the proportion between 1.0 and 2.0.

11. The mean of a normal distribution is 75 and the standard deviation is 3.
    (a) What proportion of values lies between 72 and 78?
    (b) What values are in the upper tenth of the distribution?
    (c) Approximately what proportion lies between 69 and 81?

12. Of the incomes in a normal distribution, 20 per cent are below $50 and 30 per cent are above $60. Determine the standard deviation and the mean of the distribution. (*Hint:* First express $50 and $60 as standard deviates).

13. Prepare a "less than" cumulation of suicide frequencies from Table 3.1.1d and plot the corresponding *CFP*. Using the mean of 12.5 and the *SD* of 5.4, determine (a) the proportion of suicide rates between the mean and a point one *SD* above the mean; (b) the proportion between $\bar{X}$ and a point one *SD* below; (c) the proportion between points one sigma above and below $\bar{X}$, respectively. What does this result demonstrate as regards the normality of the frequency distribution of suicide rates?

In Evansville, Indiana, there were roughly 120,000 whites and 9,000 Negroes, so that:

$$\begin{aligned} IQV &= \frac{120 \times 9}{64.5 \times 64.5} \times 100 \\ &= \frac{1{,}080}{4{,}160} \times 100 \\ &= 26\% \end{aligned}$$

Thus, as gauged by the *IQV*, there is twice as much racial heterogeneity in Indianapolis, which has a northerly location, as in the border city of Evansville.

An examination of the arithmetic of the formula will disclose that when observed frequencies are expressed as percentages, which are sometimes more convenient, exactly the same results will be obtained. Thus, for Indianapolis, the equation would read:

$$\begin{aligned} IQV &= \frac{85 \times 15}{50 \times 50} \times 100 \\ &= 51\% \end{aligned}$$

Qualitative variation, as here defined, is a strictly statistical characteristic, and should not be confused with the socio-psychological state which characterizes social disorganization, anomie, or social conflict. The degree of social disorganization may, of course, be related to the degree of statistical heterogeneity in regard to race, religion, ethnic background, or nativity, for this heterogeneity may be one of the conditioning factors in the attitudes of the population. Thus, it has been hypothesized that social tension increases as conflict groups approach equality in power, of which numerical parity is one element. The *IQV* is one tool for its measurement, and enables us to study more systematically such hypotheses.

QUESTIONS AND PROBLEMS

1. *Define the following concepts:*

   Index of Qualitative Variation
   Maximum Possible Differences

2. A given population is 50 per cent male and 50 per cent female, and 70 per cent white and 30 per cent Negro. Is it possible to represent both variables by a *single IQV? Comment.*

3. Is the *IQV* a standardized measure? Explain.

4. Is it possible for a group of persons to have more than one *IQV*? Explain.

5. At one college, 80 per cent of the students are men; at another, 67 per cent. Compute and compare *IQV*'s.

differences. This maximum number occurs when all of the frequencies of the individual attributes in the set are equal. Therefore the hypothetical maximum may be obtained by equalizing the frequencies (i.e., obtaining the mean frequency), pairing frequencies and computing products, and then obtaining the sum of all such products. In short, we (1) find the mean frequency, (2) square this result, and (3) take this square as many times as there are possible pairs of attributes. In the aforementioned example of nine boys and three girls, the maximum possible number of sex differences in a group of 12 would be 6 (boys) × 6 (girls) = 36, or, in this special case, the mean frequency multiplied by itself.

The relative amount of variation may now be measured by the ratio between the observed number of differences and the hypothetical maximum, expressed as a percentage:

$$\text{Index of Qualitative Variation} = \frac{\text{Total Observed Differences}}{\text{Maximum Possible Differences}} \times 100 \qquad (6.5.1)$$

In the above instance of 9 boys and 3 girls:

$$\tfrac{27}{36} \times 100 = 75\%$$

Among the 15 members of the three religious groups alluded to above, the mean number of members is, of course, five. Multiplying each "5" by every other "5" and summing these three products, we find the maximum number of differences to be 75. The observed differences, as already calculated, equal 74. Hence,

$$IQV = \tfrac{74}{75} \times 100 = 99\%$$

This index will always vary between zero and unity. If the numerator is zero, the index will likewise be zero, and will reflect the complete absence of variation. In the event of an equal division of observed frequencies of attributes, the numerator and denominator will be identical, and the index will be 100 *per cent, reflecting maximum heterogeneity, or variation.* Intermediate degrees of hetergeneity will take on intermediate index values.

Use of the Index. This index can be used to compare, for example, the relative amount of racial homogeneity in two or more communities. In Indianapolis there were in 1950 approximately 363,000 whites and 64,000 Negroes. Therefore:

$$\begin{aligned} IQV &= \frac{363 \times 64}{213.5 \times 213.5} \times 100 \\ &= \frac{23{,}232}{45{,}582} \times 100 \\ &= 51\% \end{aligned}$$

Norming Operations 7

Section One

Elementary Norming

The Need for Norming. A person does not ordinarily respond to an event as a raw, isolated fact, but reacts to it rather in terms of some norm, whether expressed or implied, in his social background. The individual observer thereby supplies the mental backdrop against which the event is judged or interpreted. For example, an annual income of $5,000 is not perceived as a detached item, but rather in relation to a standard based on the experience of the observer; 100 births in a community in a given year have meaning only in terms of a comparison with such relevant data as the size of the population, the number of births of the preceding year, or the number of births in comparable communities. If the standard against which the comparison is made is not expressly stated, the observer will unwittingly provide one himself in accordance with his own previous experience. There is, of course, the possibility that the individual will do this inexpertly, or will use some norm other than the one intended by the investigator who compiled the data, and thus may extract meanings from them that are inappropriate.

The difficulties will be still greater when comparisons are made between two or more sets of values based on populations that are unlike in significant respects. For example, a comparison of the intelligence of Negroes and whites in the United States on the basis of mental tests would be spurious and misleading, since performance on such tests is known to improve with level of education, which is higher among whites. Likewise, an evaluation of two or more colleges, based solely on a comparison of the average incomes of the respective graduates, would be of doubtful validity. A person's income is affected not only by the quality of his college training, but

6. A population is distributed into four ethnic groups as follows:

German	60%
French	20
Swedish	15
Irish	5
	100%

Calculate the *IQV*.

SELECTED REFERENCES

McCarthy, Philip J., *Introduction to Statistical Reasoning.* McGraw-Hill Book Company, Inc., New York, 1957. Chapter 5.

Moroney, M. J., *Fact from Figures.* Penguin Books, Baltimore, 1954. Chapter 5.

Yule, G. Udny, and M. G. Kendall, *An Introduction to the Theory of Statistics,* 14th edition. Hafner Publishing Co., 1950. Chapter 6.

cephalic index is a whole-whole ratio between two cranial measures — breadth and length — for the purpose of distinguishing quantitatively the round and oval head shapes. Such ratios are multiplied by 100 to clear the decimal point.

$$\text{Cephalic Index} = \frac{\text{Width}}{\text{Length}} \times 100$$

The familiar *Intelligence Quotient* (IQ) is a ratio between the mental and the chronological age of the tested person, which permits persons of various ages to be compared on intelligenee. Thus:

$$\text{IQ} = \frac{\text{Mental Age}}{\text{Chronological Age}} \times 100$$

It is 100 when the chronological age and mental age are exactly equal. Other common ratios employed in socio-economic analysis include: person-room, population-land, and child-adult ratio.

Arithmetically speaking, *ratio* is a generic term, indicating a comparison by division, implied or calculated. Therefore, any person may construct any ratio his needs may suggest. However, the above examples are distinguishable by their conventionality and practicality; these quotients as well as many others are widely understood and employed. They are not capricious comparisons for personal use; they are established devices which are used for the transmission of information, and they therefore take on the characteristic of a socially accepted norm.

Rates. A rate is essentially an arithmetic mean. It is the average number of occurrences of one variable per unit of another. Thus, a rate of 20 miles per gallon is the mean consumption of fuel in which the mileage would almost certainly vary from one gallon to another. The "batting average" could with equal logic be labeled a "batting rate" if it were not for the inertia of linguistic habits. A batting average of .300 could be interpreted to mean three-tenths of a hit per trip to the plate, 30 hits per one hundred trips, or 300 hits per 1,000. Since all rates are based on past observations, they provide a theoretical expectation for the future. Hence, certain calculated rates and averages are sometimes referred to as *expected* values.

Being essentially a condensation of a large number of observations, the rate has proved to be an effective measure for social analysis. A large number of rates have become conventional in the field of sociology: the marriage rate, the crime rate, birth and death rates, and innumerable modifications of these rates, which are tools in every sociological workshop.

Basically, a rate is a statistical compound of two variables: the *problem variable* and the *norming variable.* In the construction of rates, the selection of the norming variable, to which the problem variable will be related, is also by his ability, and "connections"—traits that are not uniformly distributed among college populations.

One of the important functions of the statistical method is to furnish the techniques by which single quantities may be properly interpreted, or two or more meaningfully compared. There are many devices by which these objectives are accomplished, several of which are assembled and discussed in this chapter. Collectively, they belong to a family of procedures which we may call *norming operations*, since they set up an appropriate statistical norm in terms of which the raw data are expressed and thereby rendered comparable.

The process of norming is already familiar to the student from his experience with standard (sigma) deviates, whereby diverse variables (income, age, and so on) are made comparable by expressing each as a relative distance from its own mean. Similarly, the coefficient of relative variation renders two or more standard deviations comparable to one another by expressing each as a percentage of its mean. Here we shall analyze certain other forms of this important type of statistical reasoning as they apply to sociological materials. In rough order of complexity, we may norm by: (1) percentages, (2) ratios, (3) rates, (4) indexes, (5) subclassification, and (6) standardization.

Percentages. The simplest form of norming is the reduction of a series of absolute figures to a standard numerical base compatible with the habits of thought established by our decimal system. Cumbersome and confusing absolute figures are therefore customarily expressed as so many per hundred ("per cent"), per thousand ("per mill"), per million, or in some other multiple of ten. Instead of quoting the enrollment of men and women college students as 9,244 and 4,622, we convert these values into 66.7 and 33 3 per cent respectively. So universal has this elementary practice become, that the fundamental principle on which it rests is not fully appreciated by the layman unless circumstance compels him to examine the logical foundations of his habits, as when the traveler is forced to convert miles to kilometers, dollars to pounds, meters to yards.

Ratios. Two values may be compared in ratio form, or the one expressed as a *multiple of the other*. Ratios vary in composition: they may be part-part ratios of frequencies within the same set, or whole-whole ratios between the frequencies of two selected variables.

Thus, the *sex ratio* may be viewed as a part-part ratio; it compares the number of males in the population to the number of females. In 1950 there were 74,200,085 males in the United States, and 76,016,025 females. The ratio between these bulky numbers is much more easily retained when it is expressed in the form of 97 males to 100 females, or simply 97, as it would be conventionally quoted. In the field of anthropometry, the should logically consist of those individuals that could possess the problem attribute — collectively known as the "exposed" group. While death may occur to anyone in the population, births, marriage, and divorce may not. Hence, a rate based on a judiciously selected exposed population is less subject to distortion by extraneous factors. Birth rates, marriage rates, and divorce rates calculated on a total, unselected population are therefore usually referred to as a "crude" rate; others are labeled "specific" or "refined." The advantage of the crude rate lies in its ready comprehensibility, the economy of tradition, and its utility for rough unspecialized purposes. The advantage of the specific rate lies in its precision and its serviceability in technical and professional research.

The Index. In statistics, the index, a term which is colloquially as well as technically used for various types of measures, usually pertains to relatively complicated ratios or sets of ratios. As a derived measure, it is designed to express simply the variation in a given set of values which in the raw form would be quite unintelligible. In its more formalized version, it usually refers to a ratio between two values, one of which is taken as the norm, or an expectation, against which the other is measured.

Thus, the cost-of-living index compares the average of prices in a particular year with the average for the "normal" year. An index of 130 in 1955, on a base year of 1949, indicates that the cost of living has increased 30 per cent over the base year taken as 100. While such a deceptively simple index may be glibly quoted by every columnist, nevertheless its internal composition in coverage, weighting, method of averaging of observations — as well as the choice of the base period — all testify to its statistical complexity. Analogously, the *IQV* (Chapter 6) compares the observed number of differences among attributes in a given set with the maximum possible number of differences, which in this case serves as a norm.

Table 7.1.1 illustrates the construction and use of an index for the measurement of the social stratification of university students, and the under- and over-representation of the various social classes. The logic of the indexes constructed in this tabulation is as follows: daughters of professional parents comprise 19.3 per cent of the total female student body, whereas in the state of Indiana, the professional group itself comprises only 4.7 per cent. The relationship between these paired percentages would measure the differential class opportunity. There are two procedures by which we could measure the differentials: (1) by the simple discrepancies in percentage points; or (2) by the normed indexes.

In respect to the first alternative, an examination of the discrepancies reveals a progressive change from excess to deficit, as we proceed down the social scale. However, these simple differences are absolute differences, in that they are not normed for magnitude or origin. Just as a loss of $5 from $1,000 represents a smaller *relative* discrepancy than a loss of $5 from critically important. For example, in computing the birth rate, it is necessary to decide on the choice of the norming population to which the absolute number of births is to be compared. This could be the total population, the number of women of child-bearing age, or the number of married women of child-bearing age. The most commonly quoted birth rate — though it is not the most discriminating — is the crude rate based on the total population: men, women, and children.

A second consideration in the construction of a rate involves the choice of a standard numerical base: 10, 100, 1,000 or multiples thereof. The function of the numerical base is merely to clear the decimal places for tabular convenience and facilitate quotation and quick understanding. However trivial this operation may seem, the numerical base is often fixed by convention, and one has no choice other than to comply with prevailing usage. Thus, a birth rate, quoted as 24, would be internationally understood to mean 24 births per 1,000 general population in a given year and territory. It is more readily perceived, and more securely retained, than 768 out of 32,462. Its calculation would have been as follows:

$$\begin{aligned} \text{Number of births} &= 768 \\ \text{Total population} &= 32{,}462 \\ \text{Numerical base} &= 1{,}000 \\ \text{Birth rate} &= \frac{768}{32{,}462} \times 1{,}000 = 24 \end{aligned}$$

The generalized formula would therefore read:

$$\text{Rate} = \frac{\text{Frequency of Problem Variable}}{\text{Frequency of Norming Variable}} \times \text{Numerical Base}$$

In notation:

$$\text{Rate} = \frac{PF}{NF} \times NB \tag{7.1.1}$$

$$\begin{aligned} \text{where } PF &= \text{problem frequency} \\ NF &= \text{norming frequency} \\ NB &= \text{numerical base} \end{aligned}$$

Occasionally, neither the norming variable nor the numerical base is as solidly founded in convention as is the case with general birth and death rates. In such instances, a certain discretion is permitted, the outcome of which must then be specified by the compiler. Divorce rates may be calculated on the general population, or the number of marriages in the same year and area, or even on the number of marriages during the preceding ten years, which account for most of the divorces. Delinquency rates may be calculated for specific age and sex groups; marriage rates on the age group of 14 years and over.

From previous illustrations, we may discern that the norming variable of the series. Assume, for example, that the mean death rate of a set of cities is 10.5. If all factors operating in the production of death rates in the respective cities were uniform, all the death rates would of course be identical, and therefore equal to the mean of the series. We state that the mean is the expected, or theoretical, value under these conditions. Hence, to measure the force of the differential factors, the individual rates are measured against the mean. For example, if a city has an observed death rate of 7, we would compute the index as follows:

$$\begin{aligned}\text{Index} &= \frac{\text{Observed Rate}}{\text{Expected Rate}} \times 100 \qquad (7.1.2)\\ &= \frac{7}{10.5} \times 100\\ &= 67\%\end{aligned}$$

This means that the given city has a death rate that is 67 per cent of the grand average. By this technique, any city could at once locate itself on the scale relative to the general mean. This device of norming is analogous to the *CRV* in that it expresses the raw value as a multiple of its own mean.

The *seasonal index*, which is not elaborated here, is similarly calculated as a ratio between a given monthly measure and the average measure for the twelve-month series, and is used to measure the fluctuations in births, deaths, production, and certain other economic indices.

Questions and Problems

1. Define the following concepts:
   Norming Operation
   Ratio
   Percentsge
   Rate
   Problem Variable
   Norming Variable
   Numerical Base
   Exposed Group
   Index
   Expected Value

2. In 1950, Indianapolis had a population of about 400,000, and the number of births in that year was 7,672. What was the birth rate per 1,000?

3. In the eighteenth century, European churches maintained records of births and deaths, which were later used by population scholars to estimate the sizes of cities. Thus, in Berlin, about 1700, there were 178 deaths. The estimated ratio of deaths to total population was 1:35. Calculate the death rate per 1,000. Estimate the size of the city of Berlin of that period.

4. Between 1925–1935, the birth rate dropped from 25 to 19. Calculate the permillage-point decrease (the decrease per 1,000); the percentage decrease.

Table 7.1.1 *Computation of Index, Socio-Economic Classification, Women Students and State Population, Indiana University (1950) and State of Indiana (1940)*

Socio-Economic Class	Indiana University		Indiana 1940	Diff. % Points	Index
	N	*Per Cent*	*Per Cent*		
Professional .	408	19.3%	4.7%	14.6	411
Proprietors, Managers	507	24.0	4.7	19.3	511
Dealers . .	290	13.7	5.0	8.7	274
Clerks	286	13.5	12.8	.7	105
Farmers . . .	196	9.3	15.9	−6.6	58
Skilled and Foremen .	267	12.6	17.0	−4.4	74
Semi-skilled . .	94	4.5	19.9	−15.4	23
Unskilled . .	65	3.1	20.0	−16.9	15
Total. .	2,113	100.0%	100.0%		

Source: Kate and John H. Mueller, "Class Structure and Academic Success," *Educational and Psychological Measurement*, XIII, 1953, p. 468.

$10, so the differences between the percentage points must not be interpreted uniformly without reference to the bases from which they are calculated.

This need for a normed evaluation is satisfied by the index, the second proposed alternative. Thus, if attendance were evenly distributed among all the classes of the population, one would of course "expect" the professional group, which amounts to 4.7 per cent of the general population, to supply 4.7 per cent of the student body. The *expected* proportion of students and the *observed* proportion would in that case be identical, and the ratio between them would be unity, and set equal to 100. In actuality, the professional segment of the university student body is 19.3 per cent, which is $\frac{19.3}{4.7} \times 100 = 411$ per cent of the corresponding state percentage, or 4.11 times the expected value. In like manner, we norm all other socio-economic percentages on their own respective proportion. A glance down the column of indexes will immediately and accurately attest to the progressively declining representation of the respective social classes as we proceed from high to low. This is an effective consolidation of a large and somewhat complicated array of data in the general field of social stratification.

Another approach to the formulation of an index would be the norming of a series of values on the mean of the series. The index, in that case, would be simply computed as the ratio between a given value and the mean of the series. Assume, for example, that the mean death rate of a set of cities is 10.5. If all factors operating in the production of death rates in the respective cities were uniform, all the death rates would of course he identical, and therefore equal to the mean of the series. We state that the mean is the expected, or theoretical, value under these conditions. Hence, to measure the force of the differential factors, the individual rates are measured against the mean. For example, if a city has an ohserved death rate of 7, we would compute the index as follows:

$$\begin{aligned}\text{Index} &= \frac{\text{Observed Rate}}{\text{Expected Rate}} \times 100 \qquad (7.1.2)\\ &= \frac{7}{10.5} \times 100\\ &= 67\%\end{aligned}$$

This means that the given city has a death rate that is 67 per cent of the grand average. By this technique, any city could at once locate itself on the scale relative to the general mean. This device of norming is analogous to the *CRV* in that it expresses the raw value as a multiple of its own mean.

The *seasonal index*, which is not elaborated here, is similarly calculated as a ratio between a given monthly measure and the average measure for the twelve-month series, and is used to measure the fluctuations in births, deaths, production, and certain other economic indices.

Questions and Problems

1. Define the following concepts:
   Norming Operation
   Ratio
   Percentage
   Rate
   Problem Variable
   Norming Variable
   Numerical Base
   Exposed Group
   Index
   Expected Value
2. In 1950, Indianapolis had a population of about 400,000, and the number of births in that year was 7,672. What was the birth rate per 1,000?
3. In the eighteenth century, European churches maintained records of births and deaths, which were later used by population scholars to estimate the sizes of cities. Thus, in Berlin, about 1700, there were 178 deaths. The estimated ratio of deaths to total population was 1:35. Calculate the death rate per 1,000. Estimate the size of the city of Berlin of that period.
4. Between 1925–1935, the birth rate dropped from 25 to 19. Calculate the permillage-point decrease (the decrease per 1,000); the percentage decrease.

Table 7.1.1 *Computation of Index, Socio-Economic Classification, Women Students and State Population, Indiana University (1950) and State of Indiana (1940)*

SOCIO-ECONOMIC CLASS	INDIANA UNIVERSITY		INDIANA 1940	DIFF. % POINTS	INDEX
	N	*Per Cent*	*Per Cent*		
Professional	408	19.3%	4.7%	14 6	411
Proprietors, Managers .	507	24.0	4.7	19.3	511
Dealers .	290	13.7	5.0	8.7	274
Clerks	286	13.5	12.8	.7	105
Farmers . .	196	9.3	15.9	−6.6	58
Skilled and Foremen. ..	267	12.6	17.0	−4.4	74
Semi-skilled . . .	94	4.5	19.9	−15.4	23
Unskilled	65	3.1	20.0	−16.9	15
TOTAL . .	2,113	100.0%	100 0%		

Source Kate and John H. Mueller, "Class Structure and Academic Success," *Educational and Psychological Measurement*, XIII, 1953, *p.* 488.

$10, so the differences between the percentage points must not be interpreted uniformly without reference to the bases from which they are calculated.

This need for a normed evaluation is satisfied by the index, the second proposed alternative. Thus, if attendance were evenly distributed among all the classes of the population, one would of course "expect" the professional group, which amounts to 4.7 per cent of the general population, to supply 4.7 per cent of the student body. The *expected* proportion of students and the *observed* proportion would in that case be identical, and the ratio between them would be unity, and set equal to 100. In actuality, the professional segment of the university student body is 19.3 per cent, which is $\frac{19.3}{4.7} \times 100 = 411$ per cent of the corresponding state percentage, or 4.11 times the expected value. In like manner, we norm all other socio-economic percentages on their own respective proportion. A glance down the column of indexes will immediately and accurately attest to the progressively declining representation of the respective social classes as we proceed from high to low. This is an effective consolidation of a large and somewhat complicated array of data in the general field of social stratification.

Another approach to the formulation of an index would be the norming of a series of values on the mean of the series. The index, in that case, would be simply computed as the ratio between a given value and the mean

5. In a city of 500,000, the sex ratio is 92. Calculate the number of males.

6. If a college has an enrollment of 1,236, of which 692 are males, what is the sex ratio for the school? Explain briefly what this ratio means.

7. Suicide rates are calculated on a numerical base of 100,000. In the years 1948–1952 there were 1,055 attempted suicides in Seattle (King County); of these 468 were completed, males completing 353 and females 115. The sex ratio is 100 in Seattle's population of 732,992. What is the annual suicide rate of each of the following: males, females, and total? Calculate the ratio of attempted to successful suicides.

8. In 1939, coffee cost 22.3 cents per pound. In 1949, coffee was 52.4 cents per pound, and in 1957 its price had risen to 79 cents per pound. If 1939 is considered the base year, what would be the price index for 1949 and 1957? Interpret your answer. (Source: U.S. Bureau of Labor Statistics)

9. The sex ratio in City A is 100, in City B, 50. The total population in the two cities is identical. Does City A have twice as many men as City B? Explain.

10. (a) Compute indexes of Negro and white participation in relief for each city. (Table 7.1.2)
    (b) Array Negro indexes for Northern and Southern cities separately; calculate median of each array and compare.

SECTION TWO

Subclassification and Standardization

Norming by Subclassification. Just as a single absolute value is devoid of meaning unless related to an appropriate norm, likewise a single rate or percentage possesses little if any sociological meaning when standing alone. It acquires meaning only when placed side by side with analogous rates for purposes of comparison. Accordingly, we compare the birth rate of one state with that of another, the suicide rates of Negroes and whites, the marriage rates of Catholics and Protestants. Such comparisons presumably reflect differences in fertility by state, in propensity to suicide by race, and in proneness to marriage by religion. However, the natural inference that there is an association, or even a cause-and-effect relation, between such paired variables should not be too hastily made, because the observed variation among such rates may be due at least in part to the operation of factors other than those explicitly identified in the classification, but which still influence the events under observation. Such factors are therefore labeled *concealed* factors, for they do their work invisibly in the tabulation before us. Many comparisons between two or more sets of observations are disturbed by the operation of such concealed factors, which distort the differences and lead to false inferences. Thus, the higher birth rate in

Table 7.1.2 *Families on Relief, White and Negro, Selected Cities of U.S., 1934*

City and State	Relief Population 1934			Census 1930		
	Number	*% White*	*% Negro*	*Number*	*% White*	*% Negro*
Akron, Ohio.	8,565	80.2	19.7	62,557	96.0	3 9
Ansonia, Conn.	632	83.1	16.9	4,602	94.1	5 8
Atlanta, Ga.	18,718	38.5	61.5	67,749	65.4	34.6
Baltimore, Md	40,880	56.1	43.7	193,991	82.8	17.1
Benton Harbor, Mich.	819	80.5	19.4	4,133	93.8	6.0
Biloxi, Miss.	918	70.2	29.7	3,645	80.2	19.7
Birmingham, Ala.	15,813	37.9	62.1	64,263	58.9	41.1
Bowling Green, Ky...	272	75.0	25.0	3,332	78.2	21.8
Charleston, S.C...	4,715	46.4	53.6	16,898	48.8	53.1
Charlotte, N.C. ...	2,525	28.8	71.2	19,243	68.7	33.3
Chicago, Ill.. ...	122,140	75.9	22.9	842,578	92.0	6.5
Cincinnati, Ohio	19,460	60.7	39.3	122,511	89.8	10.3
Cleveland, Ohio	46,144	75.7	24.2	221,502	91.9	7.9
Detroit, Mich.	31,370	74.1	25.2	370,293	92.6	6.9
Evansville, Ind.	4,517	77.2	22.8	25,718	93.3	6.7
Gastonia, N.C.. ..	289	72.8	27.7	3,697	78.2	21.8
Houston, Tex.	12,229	50.7	39.6	75,408	73.7	22.5
Indianapolis, Ind.	16,666	66.2	33.7	98,610	87.8	12.1
Jackson, Miss.	2,420	37.5	62.4	11,065	56.6	43.4
Kansas City, Mo.	13,132	69.9	29.3	108,641	88.9	10.5
Lake Charles, La.	815	34.6	65.4	3,884	61.6	33 3
Lakeland, Fla.	1,233	55.1	44.9	5,040	78.7	21.3
Lexington, Ky..	1,654	43.0	57.0	12,026	67.7	32.3
Little Rock, Ark....	3,670	50.2	49.8	20,026	74.3	25.7
Los Angeles, Calif....	57,960	76.4	11.7	368,508	90.3	3 0
New Orleans, La.....	14,812	34.9	65.0	111,936	68.9	30.8
New York, N.Y. ..	272,880	84.9	14.8	1,722,954	95.3	4.5
Norfolk, Va........	3,750	20.4	79.6	31,859	63.7	36.0
Pittsburgh, Pa.......	44,996	76.3	23.6	155,079	91.6	8.3
Urban United States	2,019,940	78.6	18.9	17,372,524	91.3	7.6

Source. Original data, unpublished.

one state may be due not to the greater fertility of its population, but rather to its larger percentage of women of child-bearing ages, an adventitious factor which is not manifest in the quoted crude rate.

An illustrative test of that possibility is provided by an analysis of the crude over-all birth rates of metropolitan California and rural Kansas, each quoted at 14.8 in 1940. At first glance, this identity of rates is rather startling, since rural populations are known to be more fertile than urban areas We may suspect that the population of California has a significantly larger proportion of women in the child-bearing ages than that of Kansas. In order to determine preliminarily whether the concealed factor of age has contributed to this strange outcome, we would have to prepare a table in which men are excluded and the age factor is identified. We would therefore: (1) stratify the respective populations by sex and age to determine the proportion of women of child-bearing age; (2) subclassify these women by age in appropriately small intervals; and (3) compute the age-specific birth rates for these intervals. Stratification reveals that the proportions of child-bearing women in California and Kansas are 28.8 per cent and 24.1 per cent respectively. The age-specific birth rates of these groups turn out to be as shown in Table 7.2.1. With one insignificant exception (age group 15–19), the age-specific birth rates of rural women are in fact higher than those of metropolitan centers, thereby sustaining the hypothesis which the subclassification was undertaken to test. It was these differences which were concealed in the crude rates, and which this procedure has exposed. These lower age-specific birth rates in California are, however, compensated for by an appreciably larger proportion of women of child-bearing ages, which then explains the equality of the crude rates between urban California and rural Kansas. The women of California are less fertile but there are many more of them.

Table 7.2.1

Age-Specific Birth Rates, Rural Kansas and California Cities of 100,000 and Over, 1940

Age Group	Age-Specific Birth Rate	
	Urban California	*Rural Kansas*
15–19	42.6	32.9
20–24	120.2	127.1
25–29	99.8	126.0
30–34	58.6	84.8
35–39	24.2	52.3
40–44	5.7	20.1
45–49	.5	2.0

By permission from Table 8.4, p. 150, *Population Problems*, 4th ed., by Warren S. Thompson. Copyright 1953. McGraw-Hill Book Company, Inc.

Norming by subclassification is frequently referred to as *holding factors constant*, and represents a familiar approach, colloquially expressed as "all other things being equal." Extraneous factors must be held constant in order to eliminate their disturbing effects. A death rate, for example, is usually interpreted as a measure of the vital health of a community. However, if a community with a relatively high death rate happens to have an older population than another community, that rate cannot be interpreted as a measure of vital health, but must be ascribed to the historical accident of an older population. Such a loading is quite irrelevant to the health conditions we wish to measure. The low death rate in the United States has not been altogether due to superior national health, but partly to the fact that the American population consists of such a large proportion of young persons who are not in the most fatal period of life.

Similarly, a comparison today between the crime rates of the native-born and the foreign-born populations in the United States (Table 7.2.2a) is disturbed by the fact that crime, like death and marriage, is age-linked. The foreign-born are concentrated in the older groups where crime is less prevalent, whereas the native-born are concentrated in the younger groups where crime is more prevalent. In 1950, the median age of the foreign-born was roughly 56, the native-born, only 31 years. If the crude rate is taken as a measure of propensity to crime, then the age distribution of the native-born, which is heavily loaded with younger persons, should not be permitted to work against them. Somehow, the differential effect of the irrelevant factor of age will have to be removed in order to permit a valid comparison between the two populations. The somewhat lengthy procedure by which such a removal is effected is set forth in the next set of tables.

Table 7.2.2a *Crude Crime Rates, Native-Born and Foreign-Born Populations, Ages 15–75*

Nativity	Population	Crimes	Crime Rate per 1,000
Native-born	30,000	125	4 2
Foreign-born	20,000	72	3 6

Source: Hypothetical

Table 7.2.2a presents the native-born as having a higher crude crime rate (4.2) than the foreign-born (3.6). It is a plausible hypothesis, however, that the higher rate for native-born is a result of their younger ages rather than a greater predisposition to crime. In order to investigate this hypothesis, it is necessary to make comparisons for each age group separately, which in turn requires that we subclassify *offenders by age;* that is to say, we norm by *subclassification.* This is done in Table 7.2.2b.

Table 7.2.2b *Age-Specific Crime Rates, Native-Born and Foreign-Born Population*

AGE	FOREIGN-BORN			NATIVE-BORN			PER CENT DIFFERENCE BETWEEN RATES
	Population	*Crimes*	*Rate per 1,000*	*Population*	*Crimes*	*Rate per 1,000*	
15–24	2,000	20	10	8,000	64	8	−20%
25–44	8,000	32	4	17,000	51	3	−25%
45–74	10,000	20	2	5,000	10	2	0%
	20,000	72	3.6	30,000	125	4.2	$+16\frac{2}{3}$%

We have now put ourselves in a position to answer the query: "Is the conclusion implied by the crude difference between rates confirmed within the more detailed categories?" We should recall that the native-born have a higher crude crime rate than the total foreign-born by .6 of a permillage point (4.2 − 3.6), or $16\frac{2}{3}$ per cent $\left(\frac{.6}{3.6} \times 100\right)$. But when comparisons are drawn between matched age categories — that is, when we norm on age — we find the relative positions of the native-born and the foreign-born reversed. Thus, our hypothesis is sustained. The native-born now show equal or lower rates in each age group; significantly, they have lower rates by 20 per cent and 25 per cent for the age groups in which crime rates are highest. Yet their over-all crude rate is higher. This *seeming inconsistency* is again explained by the differential age loadings: 27 per cent of the native-born, but only 10 per cent of the foreign-born, are under 25 years of age — the crime-bearing age. We could similarly have subclassified by sex, nationality, religion, race, or any other variable which showed promise of explaining plausibly the results in hand — always assuming, of course, that the additional data are available for such retabulations.

Norming by Standardization. It is evident from all this that one cannot generalize to the total groups from the comparison of matched subgroups. Not only are individual comparisons of unequal weight, but in addition they usually display considerable variation: observe, for example, the variation among the differences between the age-specific birth rates of rural and urban women (Table 7.2.1). A method must therefore be devised to obtain a simple, quotable, over-all refined rate, which conforms to the information yielded by the truer, specific rates, but which is free of the varied weights of the subgroups. This is the method of *standardization*, and the rate obtained will be the *standardized rate*.

The student has already inferred that the higher crime rate of the native-born must be attributed to the higher proportion of young persons in that population. It is this excessive loading of young people, rather than an excessive tendency to commit crime, that explains the higher crime rate of the native-born. This analysis of the age-specific crime rates already answers our question about the difference in rates, with age controlled.

But we do not yet have single rates for each of the two groups as convenient and terse as were the now discredited crude rates. A valid comparison between the single rates can be set up, free of the concealed effect of age — which is the devil in the works — if we could determine the number of crimes the native-born would have committed *if* they had had the same age distribution as the foreign-born. In other words, we must proceed as if the two populations had the same age distribution, in order to equalize their exposure to crime. In short, we standardize on the age factor. The statistical procedure consists in joining the age-specific crime rates of the native-born to the age distribution of the foreign-born, as shown in Table 7.2.2c. For example, where the native-born age-specific crime rate is 8 per thousand, and the foreign-born population is 2,000, the hypothetical number of native-born crimes is 16. The total of all such hypothetical frequencies is 60.

Table 7.2.2c *Standardization: Native-Born Crime Rate Standardized on Foreign-Born Age Distribution*

AGE	FOREIGN-BORN POPULATION	NATIVE-BORN CRIME RATE	CRIMES EXPECTED
15–24	2,000	8	16
25–44	8,000	3	24
45–74	10,000	2	20
	20,000	3	60

$$\text{Standardized Native-Born Rate} = \frac{60}{20{,}000} \times 1000$$
$$= 3$$

By reducing the absolute number (60) of crimes to a rate of 3 per 1,000, the native-born crime rate is standardized on the age composition of the foreign-born population. In effect, we have given to the native-born the "advantage" of the age distribution of the foreign-born. This eliminates the effect of the divergent age loadings, and thereby holds the factor of age constant.

By this process, the crime rates have been reversed. The standardized rate of the native-born is now only 3 per 1,000, or .6 of a permillage point below that of the foreign-born. The same general conclusion, a higher rate for the foreign-born, would have been reached had we standardized in reverse; the age-specific crime rate of the foreign-born would have advanced from 3 6 to 5.3, an excess of 1.1 over the native-born crude rate (4.2).

While standardization gives every appearance of being a neat and impersonal routine, there is no formula which prescribes how fine the subclassification should be, or for that matter what subclassification to employ. It therefore does not release the social analyst from critical subject-matter decisions. Thus, age could have been subclassified into more than three intervals. In fact, it could have been subdivided almost indefinitely, and the greater the number of subdivisions, the more precise the comparison. However, there are practical limits beyond which subclassification need not be extended. In the illustration just given, the very coarse division of age into three intervals was enough to establish the importance of age as a factor in the crime rate, which was the limited objective.

Outcome of Standardization. In the preceding example, the standing of the two populations was reversed when age was standardized; the native-born, showing initially a higher rate, emerged after standardization with a lower rate. But such an extreme reversal is perhaps exceptional. In the following example, which compares the proportion of American men 15 years of age and over who were classified as married in 1890 and 1950, the original difference is reduced when age is controlled, but not eliminated or reversed.

In 1890, only 54.1 per cent of American men were classified as married (Table 7.2.3a), whereas in 1950 the percentage had risen to 68 9 (Table 7.2.3b), an increase of 14.8 percentage points. We ask: Does this appreciable increase in the proportion married demonstrate an increasing propensity among American men to marry? Or, could it be due merely to a higher average age more favorable to marriage? Such a shift to a higher average age could induce a rise in the per cent married, with the basic propensity to marriage remaining unchanged.

To determine whether the per cent married would have increased from 54.1 to 68.9 had the age distribution remained unchanged, we need only

Table 7.2.3a *Percentage of Males Married in Specified Age Groups, U.S., 1890*

AGE	MALES		MALES MARRIED *	
	Number	*Per Cent*	*Number Married*	*Per Cent*
15-19	3,248,711	15.71	16,746	.5
20-24	3,104,893	15.02	585,748	18.9
25-29	2,698,311	13.05	1,421,407	52.7
30-34	2,425,664	11.73	1,728,930	71.3
35-44	3,705,648	17.92	2,997,030	80.9
45-54	2,627,024	12.71	2,213,901	84.3
55-64	1,630,373	7.89	1,342,414	82.3
65 and over	1,233,719	5.97	869,925	70.5
TOTAL	20,674,343	100.00	11,176,101	54.1

* Excludes widowed and divorced

Source: U.S. Bureau of the Census, *U.S. Census of the Population: 1950*, Vol. II, *Characteristics of the Population*, Part 1, *United States Summary*, Table 102, U.S. Government Printing Office, Washington, D.C., 1953.

Table 7.2.3b *Percentage of Males Married in Specified Age Groups, U.S. 1950*

AGE	MALES		MALES MARRIED	
	Number	*Per Cent*	*Number Married*	*Per Cent*
15-19	5,323,470	9.95	166,955	3.1
20-24	5,559,265	10.39	2,217,810	39.9
25-29	5,904,975	11.04	4,381,375	74.2
30-34	5,562,315	10.39	4,690,995	84.3
35-44	10,402,195	19.44	9,046,675	87.0
45-54	8,484,515	15.86	7,267,615	85.7
55-64	6,540,100	12.22	5,320,670	81.4
65 and over	5,734,250	10.71	3,767,300	65.7
TOTAL	53,511,085	100.00	36,859,395	68.9

Source: See Table 7.2.3a.

hold age constant by applying the age-specific married rates in 1950 to the age distribution of males in 1890 (Table 7.2.3c). The resulting standardized rate for 1950 is 62.9 per cent, which is 6.0 percentage points lower than the observed crude rate of 1950, but yet 8.8 percentage points

Table 7.2.3c *Computation of Standardized Rate: Percentage of Males Married in 1950 Standardized on 1890 Age Distribution*

AGE	NUMBER OF MALES, 1890	PER CENT MARRIED, 1950	EXPECTED NUMBER OF MARRIAGES
15–19	3,248,711	3.1	100,710
20–24	3,104,893	39.9	1,238,852
25–29	2,698,311	74.2	2,002,147
30–34	2,425,664	84.3	2,044,835
35–44	3,705,648	87.0	3,223,914
45–54	2,627,024	85.7	2,251,360
55–64	1,630,373	81.4	1,327,124
65 and over	1,233,719	65.7	810,553
TOTAL	20,674,343		12,909,495

$$\text{1950 Standardized Marriage Rate} = \frac{12,909,495}{20,674,343} = 62.9\%$$

higher than the crude rate of 1890. From this computation, we may infer that 41 per cent of the increase $\left(\frac{6.0}{14.8} \times 100\right)$ between 1890 and 1950 was due to a change in the age distribution, and 59 per cent $\left(\frac{8.8}{14.8} \times 100\right)$ was due to other, unspecified factors.

Standardization of Means. Standardization is a remarkably versatile tool not limited to simple rates and percentages. Any kind of arithmetic average may be standardized, provided, of course, the requisite data on subgroups are available. For example, let us suppose that sorority women at some college have a grade-point average of 1.98, while non-sorority women have a mean grade-point of 1.88 (Table 7.2.4a) From these averages alone, it would appear that the sorority women are scholastically superior. However, the non-sorority women might object to such an interpretation on the ground that sororities contain a larger proportion of upperclassmen, who normally receive higher grades than freshmen and sophomores. The superior grades of the sorority group may be attributable in part to the concealed factor of class composition rather than merely academic ability. If this hypothesis is plausible, why not standardize non-sorority grades on the class composition of sorority women as a test?

Pursuing this suggestion, we reweight the class-specific averages of

Table 7.2.4a *Grade-Point Averages by Class, Non-sorority and Sorority Women*

Class	Non-sorority			Sorority		
	(%)	$\bar{X}$	$f\bar{X}$	(%)	$\bar{X}$	$f\bar{X}$
Freshman	40	1.8 *	72	20	1.8	36
Sophomore	30	1.8	54	35	2.0	70
Junior	20	2.0	40	25	2.0	50
Senior	10	2.2	22	20	2.1	42
	100	1.88	188	100	1.98	198

* Based on a grading system in which grade-points per credit hour are as follows: A = 3, B = 2, C = 1, D = 0, F = −1.

Source: Hypothetical

non-sorority women by the class percentages of sorority women (Table 7.2.4b). In this reweighting, the grade average of non-sorority freshmen is multiplied by the percentage of sorority women classified as freshmen (1.8 × 20 = 36). In identical manner, we calculate the remaining products. The sum of such products, divided by 100, yields the sought-for standardized mean: the grade average of non-sorority women standardized on the class composition of sorority women.

Table 7.2.4b

Computation of Standardized Mean: Non-sorority Average Standardized on Sorority Distribution by Class

Class	Sorority (%)	Non-sorority ($\bar{X}$)	$f\bar{X}$
Freshman	20%	1.8	36
Sophomore	35	1.8	63
Junior	25	2.0	50
Senior	20	2.2	44
Total	100%	1.93	193

Standardized Grade-Point Average of Non-sorority Women $= \frac{193}{100}$ $= 1.93$

The crude mean of the non-sorority group (1.88), when standardized (1.93), is still somewhat below that of the sorority group (1.98). Thus, standardization has narrowed but not wiped out the original difference between the two groups. Even when the non-sorority women are accorded the benefit of sorority class [illegible], still do not quite attain the level of sorority houses Evidently other factors besides class composition have affected the grade averages, such as courses of study, IQ differences, and the like These traits, as well as others, could be employed for an indefinite refinement of the data, but at the same time they suggest the impracticality of carrying standardization very far. The infinite potentialities of standardization remind us once again how remote is "ultimate truth" from the data which lie before us and on which belief and action must nevertheless be based.

The Standard Million. Quite frequently, the crude rates that are to be converted into standardized rates are scattered over a large territory and cover an extended duration of time. Interstate vital statistics and international statistics involve cumbersome comparisons unless there is a generally recognized and available standard. To satisfy that need on an international basis, the Standard Million of England, 1901, was often employed by common consent at the beginning of this century. Since the age distribution of a population is one of the most distorting factors in the interpretation of vital statistics, this Standard Million was composed of the age distribution of the British population given in terms of a million population.

The U.S National Office of Vital Statistics at *present* employs, for the computation of its vital statistics, a Standard Million based on the population census of April 1, 1940 (Table 7.2.5). Such a procedure standardizes the vital rates for age, thereby rendering them comparable.

Table 7.2.5

Standard Million as Determined from the Population of the United States, Enumerated as of April 1, 1940

AGE	MILLION
Under 1 year..........	15,343
1- 4 years...........	64,718
5-14 years	170,355
15-24 years...	181,677
25-34 years.	162,066
35-44 years...........	139,237
45-54 years...........	117,811
55-64 years	80,294
65-74 years	48,426
75-84 years...........	17,303
85 years and over	2,770
ALL AGES......	1,000,000

Source U.S. Department of Health, Education, and Welfare, National Office of Vital Statistics, *Vital Statistics: Special Reports*, Vol. 49, No. 34, May 1949, U.S. Government Printing Office, Washington, D.C.

Norming of Cross-Tabulations by Subclassification. Cross-tabulations are usually set up in order to exhibit statistical association, but the face value of that association cannot be perfunctorily accepted — again because of the almost certain presence of concealed factors. In order to expose such concealed factors, cross-tabulations may also be subclassified. To illustrate such norming, we manipulate data (fictitious) on 52 community areas, cross-classified by race and by degree of delinquency (Table 7.2.6a).

Table 7.2.6a *Delinquency Rates by Race, 52 Local Communities*

Race	Delinquency Rate					
	Number			Per Cent		
	High	*Low*	*Total*	*High*	*Low*	*Total*
Negro	15	8	23	65	35	100
White	10	19	29	34	66	100
Total	25	27	52	48	52	100

Source Hypothetical

The tabulation indicates that Negro areas are more often characterized by high delinquency rates than are the white areas, thus suggesting a statistical association between race and delinquency. This association is brought out more clearly by the percentage distribution which reveals that 65 per cent of all Negro areas are in the "high delinquency" category, while only 34 per cent of the white areas are so classified. In view of this crude difference between Negro and white neighborhoods, the conclusion of an association between race and delinquency seems justified.

However, such a bald conclusion without reservations, would be unacceptable to anyone experienced in the social pathology of the city. He would point out that Negroes are concentrated in the blighted areas, where the standard of living is low, whereas whites more often reside in the more favorably situated districts that enjoy a relatively high living standard. It is reasonable therefore to ask whether the difference between Negro and white areas, as regards delinquency, would hold up if Negro and white areas were normed on the same economic level. Implementing this line of reasoning, we would subclassify areas according to standard of living, and draw comparisons within homogeneous socioeconomic subclasses (Table 7.2.6b).

Table 7.2.6b *Delinquency Rates by Race and Economic Level*

Race	Economic Level					
	High Standard			Low Standard		
	Delinquency Rate			*Delinquency Rate*		
	High	*Low*	*Total*	*High*	*Low*	*Total*
Negro	1	6	7	14	2	16
White	3	18	21	7	1	8
Total	4	24	28	21	3	24
Percentage Distribution						
Negro	14	86	100	88	12	100
White	14	86	100	88	12	100
Total	14	86	100	88	12	100

Inspecting the two *partial tables* which display the *partial associations*, we find that in both tables the association between race and delinquency *has vanished*: *of the 24 economically inferior areas, 21, or 88* per cent, are in the category of high delinquency; and this is true whether the areas are Negro or white. Of the 28 areas designated as economically superior, only 4, or 14 per cent, are in the high delinquency category; and this likewise holds regardless of race of area.

In this contrived illustration, then, the incidence of delinquency is dependent altogether on economic level, and not at all on the race variable. The analysis has therefore demonstrated that the original difference between Negro and white areas was due to the disproportionate concentration of Negroes in the economically depressed areas rather than the factor of race as such. Such an association is called *spurious*, since it is actually the effect of the concealed socio-economic factor which provides the "genuine" explanation. This distinction between spurious and genuine association is an important one, and will be discussed more extensively in Chapter 10.

The uneven loading of Negroes and whites in the two broad economic strata is further clarified by Table 7.2 6c which rearranges the partial tables. Lifting an example, we learn that 15 out of 25 high delinquency areas are Negro, and of these 15, 14 are on a low economic level.

The Function of Subclassification. Basically, subclassification is a procedure for refining comparisons. It is a device for holding a variable

Table 7.2.6c *Delinquency Rates by Race and Economic Level*

Race	Delinquency Rate					
	High			Low		
	Economic Level			*Economic Level*		
	High	*Low*	*Total*	*High*	*Low*	*Total*
Negro	1	14	15	6	2	8
White	3	7	10	18	1	19
Total	4	21	25	24	3	27

constant — a basic element in all rigorous scientific procedure. Statistics are seldom what they seem on face value, and it is usually necessary to go beneatb the surface in order to expose their full meaning. Subclassification, like anatomical dissection, is one probing tool for more thorough and penetrating statistical analysis. There are two distinctive interpretations which may be applied to the process of subclassification for the purpose of holding a factor constant. In one instance, the concealrd factor may be viewed as an extraneous and disturbing influence which should be "removed" and thus rendered ineffectual; in other instances, the concealed factor may be viewed as having causal significance which will open up a new line of explanation. In a strictly technical statistical sense, these two procedures are identical, but they have different sociological implications. Thus, in comparing the crime rates of the foreign-born and native-born, we subclassified by age in order to eliminate that adventitious factor by holding it constant; whereas we subclassified urban neighborhoods by economic level in order to determine whether the association between race and delinquency was a spurious one, attributable on deeper probing to the socio-economic factor. But whether subclassification is undertaken to eliminate an extraneous variable or to isolate a "causative" factor, the mechanics of the procedure are identical. Conversely, the conceptualization of the results is independent of the procedure, which simply consists in subclassifying according to the factor assumed to be related to the variable under study.

The employment of subclassification must of course be anticipated at the outset in the collection of the data. If, for example, it is desirable to subclassify by age, as in comparisons of birth and death rates, obviously information on age must have been previously collected. If the subclassification of urban areas by economic level is contemplated, information on the economic characteristics of each area, as well as an ample supply of data, must be provided for. Such breakdowns cannot be undertaken as an afterthought. This further points up the necessity of a research design to guide the collection of the data, without which the assembled information cannot be effectively exploited.

Subclassification and Standardization Compared. Subclassification differs from standardization in that it is purely descriptive and yields as many measures as there are subclasses. A standardized rate, on the other hand, is hypothetical rather than descriptive; it is a single composite index that is weighted by a set of subclass frequencies which are used as a standard. It hypothesizes what the measure would be if the selected subclass frequencies did in fact obtain. Thus, we calculated the percentage of men married in 1950 as if they had had the same age distribution as males in 1890.

Since the standardized rate is a hypothetical rather than an observed measure, a certain amount of caution is always necessary in its interpretation. It must be remembered that standardization is basically a process of reweighting, the logic of which must always be appraised in relation to the standardized results. But all normed measures are derivative rather than raw measures. To interpret them will always require a reserve of statistical and subject-matter sophistication, the more so as the chain of derivations increases in length and complexity.

Questions and Problems

1. Define the following concepts·
   Subclassification
   Standardization
   Concealed Factor
   Spurious Factor
   Causal Factor
   Partial Association
   Partial Table

2. In 1900, nearly 100,000 children graduated from United States high schools; in 1950, over 1,000,000 graduated from high schools. Criticize the statement: "High school graduates increased tenfold in 50 years."

3. For some years the crime rate for foreign-born has been declining as compared to the rate for native-born. In view of your familiarity with immigration history, how would you, in part, explain this decline? What base would you use to calculate the rates in order to produce a reliable impression?

4. Do standardized rates necessarily differ from crude rates?

5. Explain why the effect of standardization varies according to the choice of standard population.

6. Convert the cell frequencies of Table 7.2.6a to percentages of *column* totals. Is the association between the two variables thereby modified?

7. Standardize the 1890 marriage rate of males on the 1950 age distribution (Tables 7.2.3a and 7.2.3b).

8. Using Table 7.2.2b, standardize the foreign-born crime rate on the native-born age distribution.

9. Standardize the Kansas death rate on the Michigan age distribution (Table 7.2.7). Interpret.

Table 7.2.7 *Age-Specific Death Rates, Kansas and Michigan, 1950*

AGE	KANSAS		MICHIGAN	
	Population	*Age-Specific Death Rate*	*Population*	*Age-Specific Death Rate*
0– 4	199,882	6.7	703,861	6.0
5–14	299,267	.6	1,040,657	.6
15–24	277,504	1.3	922,273	1.2
25–34	282,155	1.6	1,042,819	1.1
35–44	253,327	4.6	897,921	3.3
45–54	216,097	6.4	736,124	8.3
55–64	182,249	14.6	566,461	19.1
65–74	127,010	35.4	319,653	41.6
75–84	56,438	91.0	120,310	96.3
85+	10,770	214.2	21,687	211.6
ALL AGES	1,905,299	10.0	6,371,766	9.0

Sources: U.S. National Office of Vital Statistics, *Vital Statistics of the U.S., 1950*, Vol. III, 1953, Table 47, U.S. Government Printing Office, Washington, D.C., 1953; U.S. Bureau of the Census, *Statistical Abstract of the U.S., 1950*, U.S. Government Printing Office, Washington, D.C., 1951.

SELECTED REFERENCES

Bureau of the Census, *Handbook of Statistical Methods for Demographers*. U.S. Government Printing Office, Washington, D.C., 1951. Chapter 3.

Merton, Robert K., and Paul F. Lazarsfeld, *Continuities in Social Research*. The Free Press, Glencoe, Illinois, 1950. Pages 133–167.

Silberman, Leo, "Essential Statistical Concepts and Methods in Demography," an appendix to *Population Problems*, 2d edition, by Paul H. Landis and Paul K. Hatt. American Book Company, New York, 1954.

Zeisel, Hans, *Say It with Figures*, 4th edition. Harper & Brothers, New York, 1957. Chapters 8 and 9.

Probability 8

Section One

The Nature of Probability

The Probabilistic View of Nature. Society, as a collectivity of individuels, could not exist without more or less uniform patterns of social behavior which assure a certain fulfillment of our mutual expectations. Such assurance is equelly necessary in our relation to the behavior of physical nature.

On the other band, experience has also taugbt us that the fulfillment of our anticipations is accompanied by considerable uncertainty. It is this uncertainty which gives rise to statements of probability. We may estimate the probability of war, of an economic recession, of the opportunities for a job with or witbout an ecademic degree. We say that rain is very probable, and carry an umbrella (or not) according to our confidence in such a forecast; we may gamble on a head or a tail; we speculate on tbe sex of a new baby. Sociologists attempt to measure the probability of job success, of a juvenile delinquent's becoming an adult criminal, or of a marriage terminating in divorce. Thus, the layman and the actuary, each in his own way, estimate the probability of death before 70, of a wife's survival, and even the probabilty of twins. On such estimates, man makes numerous decisions which guide his action (he gets married or gives up smoking) and even constructs a philosophy of life (he becomes a puritan or an agnostic).

The forecast of any outcome must necessarily be based on past observations. All statements of probability are a leap into the future from the springboard of past experience. Only a prophet is emancipated from the continuity of natural events. To the ordinary mortal is not given the power of prophecy; he can only extrapolate from the past.

On the prospects of war, adequate information is difficult to come by, but the prospects of death before the age of 70 are well codified in our actuarial tables.

From all this, it is apparent that probability is not an absolute condition of nature, but rather an organization imposed on nature by the observer. It is a purely human estimation of the likelihood of an event *in the future, based on acquired knowledge.* A statement of probability will therefore have no immutable value, but rather will vary according to (a) the information held by the individual who formulates it and (b) the specific conditions under which the event is perceived. Thus, a farmer who views the cloud formation, a grandfather who contemplates the fluctuating pain of rheumatism, a meteorologist with his charts—all would forecast the weather with varying degrees of accuracy. Obviously, so far as natural law is concerned, the rain will or will not fall, independently of any weather man and his elaborate probability tables. Such events were occurring before probability statistics were ever invented, and presumably will continue to occur whether the meteorologist is looking or not. There can be no probability statement independent of the human observer who formulates it.

Nor can there be a statement of probability wholly free of assumptions regarding the conditions under which the event occurs. No measure of probability is unconditional, for the simple reason that every event occurs under given circumstances which affect the degree of its probability. The self-same farmer in Kansas will have varying probabilities of death according to whether the probability is calculated on the total population of the United States, on the general population of Kansas, or only on Kansas farmers of his age. It all depends on the composition of the *set* or *subset* in which the farmer is located. But no matter how refined the measure of probability becomes, it still represents a human classification of natural events, rather than Nature herself.

Nevertheless, for all their relativity, probability statements cannot be framed arbitrarily. Their formulation is governed by fixed rules, the validity of which has been confirmed in countless problems and applications. But before turning to the simplest of these rules, collectively known as the "calculus of probabilities," it will be profitable to examine in a general way how the predicted event is statistically conceived.

Determining and Chance Factors. We conceive of any event as a juncture or resultant of innumerable factors, which may be classified into two broad categories: *determining* factors and *chance* factors.

The determining factors are those which have been isolated and to which we attribute the explanation or causation of the event. Thus, we have learned that divergent cultural backgrounds of married couples may predispose toward divorce; broken homes are associated with specific types of delinquency. The identification of determining factors is never complete, and may in fact consist only of plausible hypotheses to be cbecked. In any case, tbe focus of science is always on tbe determining factors, since they supply the basis for the understanding. and predicting of events.

Chance factors are the remaining or unknown factors affecting an event. Now, chance is not a simple concept. It represents a miscellany of factors, and is invoked to the extent that there are no *assignable causes.* In sbort, chance is a term covering our ignorance. Nevertheless, we must come to some understanding of a concept which has proved itself so indispensable in man's effort to explain nature.

For statistical purposes, chance factors are presumed to be (1) very numerous, (2) relatively minute, (3) independent of one another, and (4) largely unidentifiahle and therefore not measurable; consequently, (5) in effect, they work collectively to produce equally likely events. When we are unahle to assign a cause, we resort to chance as an "explanation" and equiprohability as a prediction. Although there is an element of chance in the production of every event, its characteristics may best be analyzed hy using games of chance as a model — situations in which chance factors hy definition operate exclusively.

Thus, the drop of tbe ace of spades into the bridge hand of North is a result of sheer chance. (1) It is the result of innumerable factors, such as the location of the card in the previous deal, the amount of shuffling, the position of the dealer, and the like. (2) Each of these factors alone is relatively impotent and is easily counteracted by other factors. (3) These factors act independently of one another: the location of the ace in the undealt pack, tbe zeal of the shuffler, the cut of tbe opponent, the location of the bañds, are all unrelated. (4) It is impossible for anyone to conceive of, or identify, even a small proportion of the factors influencing the drop of the card, much less to measure their force. (5) Since the chance factors are assumed to impinge with equal force on all 52 cards, each card is as likely to turn up as is the ace; this means that all cards presumahly will appear in the long run with equal frequency in a given hand. A professional gambler, or "magician," on the other hand, unwilling to leave the outcome to the laws of chance, will cut the deck according to his illicit knowledge of the location of tbe ace.

One must not be misled into believing that the distinction between chance and determining factors is intrinsic and in the nature of the things. Mother Nature herself would make no such distinction. There is nothing mystical about chance; chance factors are just as materially real as are determining factors. In fact, they may be viewed as unidentified "determining" factors which play an important role in the production of events. The selection of a spouse, the choice of occupation, the outcome of a football game, the fatal accident — all may have their turning point in chance factors. No event is completely accounted for without a consideration of their potency.

Nevertheless, it may be pointless and prohibitive even to attempt to analyze them individually, especially when the event is exclusively determined by chance factors. For example, the probability of a specified hand in bridge is 1 out of 635,013,559,600. How useful would it be to expend time and energy on redefining the chance factors in order to explain how such a rare and relatively unimportant event occurred in the first place? On the other hand, classical science has, as its ideal, the transformation of all chance factors into determining factors. This would permit the prediction of the specific individual event and remove the procedure from the province of statistics. But twentieth century natural science has tempered this Newtonian ideal, because of the growing awareness that certain types of phenomena are not amenable to classical analysis. The following excerpt from Einstein and Infeld * illustrates this methodological revision in the realm of physics. In principle, it could have been written by a sociologist:

> There is a vessel containing gas. In attempting to trace the motion of every particle one would have to commence by finding the initial states, that is, the initial positions and velocities of all the particles. Even if this were possible, it would take more than a human lifetime to set down the result on paper, owing to the enormous number of particles which would have to be considered. If one then tried to employ the known methods of classical mechanics for calculating the final positions of the particles, the difficulties would be insurmountable. In principle, it is possible to use the method applied for the motion of planets, but in practice this is useless and must give way to the *method of statistics*. . . . We become indifferent to the fate of the individual gas particles. Our problem is of a different nature. *For example:* we do not ask, "What is the speed of every particle at this moment?" But we may ask: "How many particles have a speed between 1000 and 1100 feet per second?" We care nothing for individuals. What we seek to determine are average values typifying the whole aggregation. It is clear that there can be some point in a statistical method of reasoning only when the system consists of a large number of individuals.
>
> By applying the statistical method we cannot foretell the behavior of an individual in a crowd. We can only foretell the chance, *the probability*, that it will behave in some particular manner. If our statistical laws tell us that one-third of the particles have a speed between 1000 and 1100 feet per second, it means that by repeating our observations for many particles, we shall really obtain this average, or in other words, that the probability of finding a particle within this limit is equal to one-third.
>
> Similarly, to know the birth rate of a great community does not mean

* Albert Einstein and Leopold Infeld, *The Evolution of Physics*, Simon and Schuster, New York, 1938, pp. 298–299.

knowing whether any particular family is blessed with a child. It means a knowledge of statistical results in which the contributing personalities play no role.

By observing the registration plates of a great many cars we can soon discover that one-third of their numbers are divisible by three. But we cannot foretell whether the car which will pass in the next moment will have this property. Statistical laws can be applied only to big aggregations, but not to their individual members.

Whatever might have been stated in the previous discussion about the uncertainties arising out of the play of chance factors, it does not argue the entire capriciousness of chance. Although chance factors cannot be used to predict the individual event, they nevertheless produce certain regularities in the behavior of masses of events which are termed the *laws of chance*. The formulation of these laws has been an accomplishment of the last two hundred years, and these have come to be the basic ingredient in what is known as *inferential statistics*.

The Possibility Set. Statistically speaking, a probability is simply the proportion of times an event is expected to happen in the long run, in a great mass of trials. But such proportions cannot be established unless all possible different outcomes have been identified. Together, these outcomes that could occur constitute the *possibility set*. The probabilities of all possible outcomes of course sum to 1.00, since it is certain that one outcome of the set will occur on a given trial. Furthermore, alternative outcomes must be mutually exclusive, and must be at least two in number, for if there were only one possible outcome, there would be no uncertainty and therefore no probability. These minimum alternatives are frequently labeled *success* and *failure*. If attention is focused on only one possible outcome — termed "success" — only two probability values need be calculated; for example, the probability of an ace and not-ace, the probability of spade and not-spade.

But in any case, it is essential to identify all possibilities that belong to the possibility set, otherwise no probability statement can be made. Consider the question: What is the probability that Mr. Jones will vote Republican? Before we can answer that question quantitatively we must first determine all possible ways of voting (i.e., the number of parties on the ballot). Without that information, no probability statement would be possible, since the probability of a given outcome can only be measured in relation to all possible outcomes. We must know the number of all political parties in order to fix the probability of any single one.

While every logical possibility must be enumerated, they still may be variously grouped according to the purposes of the user. The given possibilities may be decomposed or recombined into any desirable *subsets*. Multiple births may be broken down into twins, triplets, quadruplets, and the like, while bome runs, triples, and doubles may be grouped together as extra-base hits. Playing cards may be classified according to suit or face value; deaths may be classified according to terminal cause. But however outcomes are arranged, the set must include all possible outcomes — that is, it must be exhaustive.

Once the possibility set has been established, we then determine the relative frequency, or weight, of each possible outcome or event; for example, the expected frequency of a bead and the expected frequency of a tail; the expected frequency of a male birth and the expected frequency of a female birth, the expected frequencies of single and multiple births, or of Republican and Democratic votes. The establishment of such expected proportions may be based on one of two principles: the *a priori*, or the *empirical* principle.

A Priori Probabilities. A toss of a coin may result in one of two outcomes: head or tail. We say the probability of tossing a head is $\frac{1}{2}$ or 50 per cent; the probability of a tail, exactly the same. Similarly, we examine the 52 cards of a well-shuffled bridge deck, and conclude that each card is equally likely to be dealt. On what evidence do we weight them equally? According to one theory, the only reply is that we have no reason to conclude otherwise: we examine the coin for symmetry, and the 52 cards for perfect uniformity, and the tosser or dealer for honesty, and by a priori reasoning conclude that the evidence is sufficient to predict that each outcome is equally likely in the long run. This method, not based on experimental test or observation, has been labeled the *principle of sufficient reason.** The examination of a die would lead to the analogous conclusion that each side would be equally likely in the long run. However, an oblong eraser or pyramid-shaped paper weight would not offer intuitive evidence that each side would turn up with equal frequency in a large number of throws. There is no ready means available to ascertain a plausible ratio between the various sides on the basis of their shapes and areas. To establish the probabilities of the respective sides, it would be necessary to toss the objects a very large number of times and record the empirical results.

In fact, the critics of the a priori school of thought contend that the belief in any a priori judgment or pure reason is a form of self-deception. The judgment that the two sides of a coin, or six faces of a die, are equiprobable rests on the actual experience of previous tosses as practiced by countless youths in every generation. Hence, judgments on heads and tails are not at all a priori, but rather based on personal experience, and even traditional expectations. We are never tempted to estimate the probability of death on an a priori basis, since the simple conditions which surround the tossing of pennies and dice do not prevail in vital statistics.

* Some authors prefer the phrase "insufficient reason," which they feel is more appropriate.

In general, the a priori method of fixing probabilities becomes wholly inadequate after we leave the very limited realm of games of chance.

Empirical Probabilities. If weigbts are assigned to possibilities on the basis of empirical observations, they are known as *empirical* probabilities. As a simple practical example of the calculation of empirical probabilities we present an abridged life table of the United States (1950) and of eighteenth century Berlin (Table 8.1.1). The purpose of this record is to set up the probabilities of death for various age groups. The elementsry outcomes are, of course, life and death in a given year. The weights of these two possibilities are based on the average number of deaths over a reasonable period of observation. The average of tbe past thereby becomes the best estimate for the probability of the future. Thus if, out of a cohort of 100,000 United States births, approximately 87,000 have survived to the age of 50, we may conclude that in tbe future, as well, anyone in a similar cohort would have an 87 per cent probability of surviving to that age. In eighteenth century Berlin, the corresponding person would have had only a 31 per cent probability of surviving to age 50. A twenty-year-old youtb in the United States would have an 80 per cent probability ($\frac{75,921}{95,366}$) of reaching his sixtietb birthday. This is, of course, the type of empirical actuarial data on which life insurance premiums are computed.

Table 8.1.1

Abridged Life Table, U.S. 1950 and Germany c. 1750

Age	Number Surviving at Specified Ages	
	U.S.	*Germany*
0	100,000	100,000
10	96,177	54,000
20	95,366	49,600
50	86,591	31,300
60	75,921	22,600

Sources Johann P. Süssmilch, *Die Göttliche Ordnung*, 2d edition, Berlin, 1762, Vol. II, pp. 319–322 (adapted); and U.S. National Office of Vital Statistics, *Vital Statistics — Special Reports, Life Tables for 1949–1951*, Vol. 41, No 1. U.S. Government Printing Office, Washington, D.C., 1954, Table 2.

the like, while home runs, triples, and doubles may be grouped together as extra-base hits. Playing cards may be classified according to suit or face value; deaths may be classified according to terminal cause. But however outcomes are arranged, the set must include all possible outcomes — that is, it must be exhaustive.

Once the possibility set has been established, we then determine the relative frequency, or weight, of each possible outcome or event; for example, the expected frequency of a head and the expected frequency of a tail; the expected frequency of a male birth and the expected frequency of a female birth, the expected frequencies of single and multiple births, or of Republican and Democratic votes. The establishment of such expected proportions may be based on one of two principles: the *a priori*, or the *empirical* principle.

A Priori Probabilities. A toss of a coin may result in one of two outcomes: head or tail We say the probability of tossing a head is $\frac{1}{2}$ or 50 per cent; the probability of a tail, exactly the same. Similarly, we examine the 52 cards of a well-shuffled bridge deck, and conclude that each card is equally likely to be dealt. On what evidence do we weight them equally? According to one theory, the only reply is that we have no reason to conclude otherwise: we examine the coin for symmetry, and the 52 cards for perfect uniformity, and the tosser or dealer for honesty, and by a priori reasoning conclude that the evidence is sufficient to predict that each outcome is equally likely in the long run. This method, not based on experimental test or observation, has been labeled the *principle of sufficient reason*.* The examination of a die would lead to the analogous conclusion that each side would be equally likely in the long run. However, an oblong eraser or pyramid-shaped paper weight would not offer intuitive evidence that each side would turn up witb equal frequency in a large number of throws. There is no ready means available to ascertain a plausible ratio between the various sides on the basis of their shapes and areas. To establish the probabilities of the respective sides, it would be necessary to toss the objects a very large number of times and record the empirical results.

In fact, the critics of the a priori school of thought contend that the belief in any a priori judgment or pure reason is a form of self-deception. The judgment that the two sides of a coin, or six faces of a die, are equiprobable rests on the actual experience of previous tosses as practiced by countless youths in every generation. Hence, judgments on heads and tails are not at all a priori, but rather based on personal experience, and even traditional expectations. We are never tempted to estimate the probability of death on an a priori basis, since the simple conditions which surround the tossing of pennies and dice do not prevail in vital statistics.

* Some authors prefer the phrase "insufficient reason," which they feel is more appropriate.

later find it profitable to bet consistently on tails, and he would be right more than half of the time. In this case, he has rejected the initial probability statement and established a new one consistent with the observed frequencies. Such rejection is known as a *statistical decision.*

The same testing procedure is applied to empirical probabilities. In the case of life tables, a certain chance variation is permitted without casting doubt on the reliability of the empirically predicted death rates. But when the variation becomes too great, the actuarial tables are revised, and a new norm or model is set up for subsequent observations.

From the foregoing discussion, it is obvious that the determination of the possibilities and their respective weightings is the core of probability calculation. The most elementary rules of such calculation will now be presented.

Questions and Problems

1. Define the following concepts:
   Probability
   Chance
   Determining Factor
   Possibility Set
   Outcome
   Event
   A Priori Probability
   Empirical Probability
   Principle of Sufficient Reason
   Expected Frequency
   Law of Large Numbers
   Equiprobability

2. Discuss probability in its following aspects: subjective, statistical, relativistic.

3. List possible chance factors in choice of occupation; of spouse; of college; of residence.

4. On a True–False examination, a student marks 50 per cent of the questions correctly. Is that evidence (a) that he has answered them by chance, or (b) that he knows 50 per cent of the subject matter? How might you determine which of these alternatives is more plausible?

5. The *New York Times*, November 7, 1948, stated that 7 of 32 past presidents had died in office. Hence, Vice President Barkley would have $\frac{7}{32}$ chances to succeed President Truman in office. Comment on the validity of this statement.

6. If 8 per cent of the students in a large class are left-handed, what would this observation indicate about the probability of finding a left-handed person in another class? Discuss.

ments. These statements must be regarded as models, idealized on infinity, with which the subsequently observed finite ratios can only occasionally agree. We will only rarely observe exactly 50 heads on 100 tosses of even a perfect penny, although the probability of heads is quoted as 50–50 Such discrepancies between expected and observed frequencies will result either from (1) continued operation of chance factors which introduce an indefinite amount of variation, or from (2) the operation of changing conditions which make the initial statement of probability invalid These discrepancies ultimately lead to one of the central issues of inferential statistics (decision-making, Chapter 12), which at this time may be given only preliminary mention: how may the discrepancy between expected and observed be interpreted — as chance variation around the stated probability value, or as evidence against the initial statement of probability?

Let us suppose that, in a large number of tosses of the same coin, a gambler observes heads 40 per cent of the time, instead of the a priori 50 per cent. At this point in the game, how should he react on the next bet? He must decide whether to accept the discrepancy as mere chance *variation to be ignored*, or *reject the initial probability statement as false* and revise his expectation

His first line of reasoning might be: the *law of large numbers* prescribes that with an increasing number of tosses, the approximation to the hypothetical true proportion must become correspondingly closer. The "deficit" of heads must have resulted from excessive *runs* of tails. Hence, sooner or later the heads will have to "catch up" in order to make up the deficit. Therefore, "I will now bet on heads." However, such a judgment is inconsistent with the usual assumption in penny tossing that each outcome is independent — that is, uninfluenced by previous ones. The coin has no "memory" of previous outcomes, and therefore heads and tails are still *equiprobable* on the next toss. After all, the 50–50 division of events holds good only for an infinitely large number of throws, or trials. Any finite number of trials is not enough to equalize the outcomes. Hence, the deficit need actually never be made up. Therefore, in spite of the unusual runs of tails the bettor should not change the pattern of his strategy. He has accepted the discrepancy as due to the play of chance factors.

On the other hand, if the run of tails persists beyond the limits of reasonable tolerance, the player might well raise the question whether the orthodox hypothesis of equiprobability was valid in the first place. Perhaps the coin was not symmetrical after all; perhaps the coin favors the tail, as the empirical evidence seems to indicate. In other words, if the *discrepancy between expected and observed frequencies becomes too great*, we should question the hypothesis, and reject it accordingly. Although there is no fixed rule which could guide the observer in determining at what point the length of the run becomes suspicious, he would sooner or arbitrarily defined as "success"; if the probability of survival is in question, then life would be considered "success." An unlimited number of repeated or concurrent trials is essential to the concept of probability. Conventionally, we speak of such an infinity of trials as *conceptually repeatable trials*, or *experiments*, since infinity can exist only in the imagination.

Elementary Rules of Calculation. While all probability calculations are fundamentally alike, they differ in detail according to whether the event under consideration is (1) *a simple event*, (2) *a joint event*, or (3) *an alternative event.*

(1) *Simple Event.* A simple event is the lowest unit of observation as defined for the data under analysis. It may be the head of a coin, cancer as a cause of death among the 200 standard causes, a multiple birth, a successful parole, a divorce, or a hit in baseball.

The elementary nature of the event is not intrinsic but is rather a matter of definition to be decided by the person doing the classifying. Clearly, cancer could be profitably decomposed into a number of subclassifications according to the site of the affliction; a multiple birth may be subclassified into twins, triplets, quadruplets, and quintuplets. Any simple outcome could conceivably be broken into finer classifications if the problem in hand so dictated. Therefore, a simple event is one that is arbitrarily so considered by the observer. To calculate the probability of a simple event, we apply:

> RULE 1. The probability of a simple event is the ratio of the frequency of that event to the total frequency of all possible events in the set.

For example, in an average year, there are about 4,000,000 births $(n + m)$ in the United States, of which approximately 46,000 (n) are multiple births. Therefore, the probability of a multiple birth is:

$$\frac{n}{n+m} = \frac{46{,}000}{46{,}000 + 3{,}954{,}000}$$
$$= \frac{46{,}000}{4{,}000{,}000}$$
$$= \frac{1}{87}$$

This ratio, a partial statement of Zeleny's well-known law, has remained amazingly representative of the whole of western civilization, which consists of innumerable trials.

When the frequencies of a set of simple events are identical, Rule 1 reduces to:

7. It was reported that 15,000 intoxicated pedestrians were killed by cars, and only 5,000 intoxicated drivers were killed. This means that the probability of a pedestrian being killed is three times the probability of a driver. Comment.

8. Comment on the statement· "Improbable events are extremely common."

9. Does the principle of chance apply to such "certain outcomes" as eclipses, the rising sun, and other events which are practically certain to occur? Comment.

10 A soldier jumped into a foxhole which had just been hit by a shell. He reasoned that a shell will not, by chance, strike the same spot again. Criticize. (*Hint:* Is the chance hypothesis the only possible hypothesis?)

11. Comment on the cliché that "lightning does not strike twice in the same spot."

SECTION TWO

Probabiity Calculation

Probability as a Frequency Ratio. To ask "What is the probability of death from cancer?" is essentially the same as asking "What fraction of all deaths are caused by cancer?" Although a competent demographer would wish to specify the cohort of population to which the question referred, this would not change the fundamental nature of probability as a frequency ratio. Hence, to answer the above question we would have to express the frequency of death by cancer as a proportion of deaths by all possible causes, including cancer. About 20 per cent of all deaths now occurring in the United States are caused by cancer; therefore, the probability or expectation of death by cancer among subsequent deaths, would likewise be 20 per cent. This experience ratio of 1 in 5 trials is used to forecast the events of the long-run, on the assumption that the conditions of death remain constant. Such a ratio is conventionally expressed:

$$Pr(E) = \frac{n}{n+m} \qquad (8.2.1)$$

in which $Pr(E)$ = the probability of a given event, E
n = the frequency of occurrence of the given event
m = the frequeney of all other events in the set

Definition of a Trial. The concept of a *trial* does not necessarily limit itself to a contrived act, such as the tossing of a coin, or the trial of a batter who may or may not get a hit. It may also refer to the interrogatory attitude of the observer who awaits the outcome of a natural situation which may eventuate in one of several ways. Thus, the life or death of a given member of a population cohort is the outcome of a trial in a statistical sense. If the probability of death is to be calculated, then death is

Rule 3. The probability of two or more events occurring together, or a joint event, is the product of the probabilities of the individual events.

Thus, the probability of two fives is:

$$\begin{aligned} Pr(5 \text{ and } 5) &= Pr(5) \times Pr(5) \\ &= \frac{1}{6} \times \frac{1}{6} \\ &= \frac{1}{36} \end{aligned}$$

(3) *Alternative Event.* We may wish to broaden the definition of success by regarding any one of a number of alternative events as constituting success, such as any one of four aces. We may bet on a horse not to "win," but to win, place, or show. Statisticians usually think of these alternatives as constituting a subset of elementary possibilities. By increasing the number of acceptable possibilities in the subset, we necessarily increase the probability of success, since we merely add the probabilities of the respective outcomes. This principle is explicit in the *addition theorem:*

Rule 4. The probability that one of two or more mutually exclusive alternative events will occur is the sum of the probabilities of the individual events.

For example, by this theorem the probability of any one of the four aces would be expressed:

$$Pr(A_1 \text{ or } A_2 \text{ or } A_3 \text{ or } A_4) = \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{4}{52}$$

The addition theorem may be applied not only to simple events, as the aces in a deck of cards, but also to joint events. Thus, in the toss of two dice previously alluded to, there are six ways of obtaining a total of 7 — (1 + 6), (2 + 5), (3 + 4), (4 + 3), (5 + 2), (6 + 1) — and two ways of obtaining 11 — (6 + 5) and (5 + 6) — on the respective faces of the pair of dice. By simple addition, we learn that the probability of throwing a 7 is $\frac{6}{36}$, and the probability of an 11 is $\frac{2}{36}$. Therefore, the probability of tossing *either* a seven *or* eleven would be 8 out of 36.

Outcomes Not Mutually Exclusive. Rule 4 holds only for events that are *mutually exclusive* — that is, for outcomes that cannot occur together on a single trial. Thus, a head and a tail cannot both occur when a single coin is tossed; a birth cannot be single and multiple at the same time. Yet, in many instances, a given object may be classified or perceived in more than one way: a playing card may be read simultaneously as an ace and a spade; an individual may be characterized as both single and Catholic. Such traits are not at all mutually exclusive since they rep-

Rule 2. **The probability of a given event in a set of equally likely events is the ratio of 1 to the total number of possibilities, k.**

That is,

$$Pr(E) = \frac{1}{k} \tag{8.2.2}$$

For instance, in the toss of a coin there are two possible outcomes, assumed to be equally numerous in the long run; hence, the probability of a head is:

$$Pr(H) = \frac{1}{1+1} = \frac{1}{2}$$

(2) *Joint Event.* Events do not necessarily occur singly; two or more distinguishable events may occur together to form a *joint event,* sometimes labeled a *compound event.* The toss of two dice, for example, will always result in concurrent events such as 1 *and* 1, 1 *and* 2, · · ·, 6 *and* 6, representing the faces of the dice.

If according to Rule 2, the probability of a 5 on Die 1 is $\frac{1}{6}$, and the probability of a 5 on Die 2 is likewise $\frac{1}{6}$, then the probability of their occurring together would be $\frac{1}{36}$. This result may be clarified visually by means of the accompanying diagram which deploys the 36 possible ways in which the two dice may fall together — that is, the 36 joint outcomes. While there is only one way in which the two fives may fall together to give a probability of $\frac{1}{36}$, a 2 and 3 may also occur as 3 and 2 — in two ways — for a probability of $\frac{2}{36}$. The probabilities of other joint outcomes may similarly be read from this diagram.

Die 1	*Die 2*	*Die 1*	*Die 2*	*Die 1*	*Die 2*
	1		1		1
	2	3	2		2
1	3		3		3
	4		4		4
	5		5	5	5
	6		6		6
	1		1		1
	2		2		2
2	3	4	3	6	3
	4		4		4
	5		5		5
	6		6		6

This cumbersome procedure of listing every possible way is not necessary in practice; instead we always apply the *product theorem,* as expressed in:

Thus, in computing the probability of drawing two aces in succession from a bridge deck of 52 cards, the probability of drawing the first ace is, of course, $\frac{4}{52}$. But the drawing of one ace (without replacing it) obviously affects the probability of drawing another ace on the next trial. There being only 51 cards remaining (including three aces), the probability of the next ace is now $\frac{3}{51}$. The joint probability is the product of the individual probabilities: $\frac{4}{52} \times \frac{3}{51} = \frac{12}{2,652}$. This means that, in the long run, out of 2,652 trials consisting of two random draws without replacement, one would expect 12 outcomes of double aces.

This illustration, which is convenient because of its easy manipulation, is a formal, routine demonstration of a familiar principle. However, in the practical affairs of social life — to say nothing of pure science — it is not always so easy to detect the presence of dependence, and still more difficult to measure its degree. Dependence is often very complex — as are all social events — and accurate calculation is often impossible. Consequently, we are frequently obliged to disregard the presence of dependence, and forgo its calculation, since the needed empirical data are inaccessible.

For example, actuaries calculate the joint probability of spouses surviving to celebrate their golden wedding. If, at marriage, the bride is 21 and the groom 26, they would have to live at least to the ages of 71 and 76, respectively. Now, we know from the general life tables for the United States that the 26-year-old male has one in three chances of surviving to that age; the 21-year-old female has three out of five chances to survive for fifty years after marriage. By applying the product theorem, we find that the probability of both surviving to the required ages would be $\frac{1}{3} \times \frac{3}{5} = \frac{1}{5}$ or .20. But this probability is computed on the assumption that the survivals of the two spouses are mutually independent. It seems incredible, however, that marriage partners are chosen at random. Since it is plausible to believe that healthy spouses tend to marry healthy mates, in that respect there are selective factors in operation, as well as chance factors. Furthermore, since husband and wife share a common way of life, the health of one is likely to influence the health of the other. Hence, their survivals are not as completely independent as the above probabilistic calculation requires. Far from being independent, there is probably a rather high correlation between the survival rates of married spouses. These individual probabilities were obtained from the general life tables, simply because more accurate ones

come of A in the joint event. The reason for this omission of a traditional clause is simply that the multiplication rule for joint events, in its most general form, holds both for independent and dependent probabilities. When events are lacking in independence, then the probabilities must be duly adjusted to allow for the degree of dependency; otherwise, the product theorem will produce invalid results. Nevertheless, the product theorem does apply to conditional probabilities, provided of course they have been properly computed.

resent different dimensions of the same object. It would therefore be erroneous to apply the simple addition theorem, which holds only for mutually exclusive events, to events which may occur together. Thus, to calculate the probability of an ace *or* a spade, we must not add the probabilities of these simple events and let it go at that. The probability of an ace or a spade is *not:*

$$\frac{4}{52} + \frac{13}{52} = \frac{17}{52} \qquad \text{(Wrong!)}$$

This probability is too high, since we are erroneously combining two non-exclusive traits. When outcomes that are *not* mutually exclusive are produced, it is obvious that the alternative outcomes will, in a certain proportion of cases, occur together. Thus, the alternative events, "ace" or "spade," will occasionally both turn up together as the "ace of spades" and must be accepted as a success.

However, in adhering to the simple addition theorem (Rule 4), the ace of spades will mistakenly occur twice, whereas every other event will occur only once. But theoretically, the ace of spades should occur only as often as every other card — once in 52 times. Although the ace of spades is accepted as a success, we are not permitted to count it twice as often as its random appearance would justify. The foregoing addition rule must therefore be amended by subtracting the duplicate "occurrence," which is measured by its joint probability, or $\frac{1}{52}$. The procedure just described is now generalized as:

> RULE 5. The probability that one of two alternative events, not mutually exclusive, will occur is the sum of the individual probabilities, minus their joint probability.

According to this rule, the probability of an ace *or* spade would be:

$$\frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$$

Independent and Dependent Outcomes. By independence is meant that a given event or outcome is uninfluenced by the preceding or concurrent events. Thus, the probability of a head on a given trial is independent of the previous falls, whether head or tail. The probability of the next outcome has not changed because of the previous event.

However, when the probability of one of the outcomes in a joint event is in any way conditioned by the previous outcome, we say that the second is *dependent* on the previous event. In that case, a revised probability of the second event must be computed to take account of that degree of dependence before the multiplication rule can be invoked.*

* The above formulation of Rule 3 omits mention of the usual provision that probabilities must be independent, i e , that the outcome of B is in no way affected by the out-

channel at that time. With all his singularities, he was also a member of a class (a subset) representing a large number of similar historical instances. The question should therefore be rephrased to read: What is the probability of persons like Caesar, under similar circumstances, having gone to Britain? Although these forces do not lend themselves to quantitative measurement, an intuitive but profitable approximation of probability can be obtained. Many important human decisions, including the adoption of scientific hypotheses and legal judgments, rest on such qualitative bases rather than on methodical quantitative procedures. Some statisticians would reserve the concept "likely" for such circumstantial judgments, and employ the concept "probable" only when estimates are founded on rigorous quantitative observations.

A third possible interpretation of the foregoing statement holds that a single outcome can never set up, modify, or test a probability statement. The reason for this is simply that probabilities are established or confirmed only by the results of a large number of trials. In that sense, statistics do fail to "tell us anything about the single case." A single parole violation would not be sufficient to disprove the correctness of a probability of 80 per cent success, established by previous empirical observations, since 20 per cent of the cases are confidently expected to be failures. It is only after a large number of subsequent observations, when the constant factors have had an opportunity to manifest their force, that the validity of a probability statement can be made plausible, although never proved. We would question the hypothesis of a fair coin after a run of three or five heads, and certainly reject it after a run of perhaps twenty-five.

Questions and Problems

1. Define the following concepts:
   Trial
   Simple Event
   Joint Event
   Alternative Event
   Addition Theorem
   Product Theorem
   Dependent Event
   Independent Event
2. According to the International List, there are 200 causes of death, of which cancer is one; hence, a person has $\frac{1}{200}$ probability of dying of that disease. Comment.
3. In a group of 6 persons, there are 4 boys and 2 girls. If we selected two individuals by chance, what would be the probability of drawing a boy first and a girl second? a girl first and a boy second? two girls? two boys?
4. For the following distribution of Jewish and Protestant males and females, calculate the probability of mixed marriages and homogamous marriages,

were not available Weightings for subsets by marital state, occupations, and other relevant classifications are not provided. Hence, strictly speaking, the joint probability of $\frac{1}{5}$ is an over- or under-statement depending on the health of the cohort. Only if men and women were randomly paired would the joint probability of .20 be valid. But even if the dependent probabilities could be as easily calculated as in the case of the two aces, it is not likely that a refinement of these calculations would greatly influence the romantic mating process.

The Probability of a Single Event. Since the concept of probability assumes innumerable repeated trials, which give the law of large numbers enough room to operate in, it is quite conventionally asserted that we cannot assign a probability to the outcome of a single trial. This apparent truism is not as simple as it appears. There are three possible reactions to this ambiguous statement

In the first place, we do act on probabilities, and we commonly do apply them to single trials. Thus, "in the clutch," a pinch hitter is selected on the differential probability of his getting a hit. The manager will reason as follows: "Jones has a batting average of .300 (empirical probability); if I send him up to bat, he is more likely to succeed than Smith, who has only a .250 average." Analogously, a person who is ill calculates his own private chances for recovery and makes plans accordingly, however indifferent the insurance company may be to the identity of the individual case in its actuarial mass statistics. Thus, each person continually codifies his experiences and probabilistically prepares for the alternative outcomes in each single instance.

In the second place, a single case may be thought of as a wholly unique event. With this interpretation, it would seem doubly clear that no probability statement could be made, since there can be no long run of unique events — for that matter, no run at all. Being unique, the event could not have occurred in the past; and obviously could not occur again. A probability statement would therefore be both impossible and useless. For example, if Julius Caesar had tossed one of his coins, we could properly and reliably state the probability of his having turned up his image thereon. But we cannot with the same precision estimate the probability of his having been in England. Nevertheless, historians are constantly speculating intelligently on the "unique" actions of Caesar and many other personages in history. How can the statistical principle forbidding such speculations be reconciled with the familiar and useful deductions of thinking people?

The deception lies in the fact that, sociologically speaking, a unique event is never totally unique. Although there was only one Julius Caesar, he still possessed traits in common with other leaders of men, not only in his broad characteristics, but also in the potentialities of a trip across the

This count will depend on whether we focus on the membership of the group or the order in which the component members are placed — that is, on whether we are dealing with a *combination* or *permutation*. Statistically speaking, combination denotes the identity of the individuals that compose a group; while permutation denotes the specific order, or arrangement, in which the members may be placed.

In order to determine the number of combinations or permutations in any given set, we could of course laboriously enumerate all possible ways, as has been done for the tosses of two dice (Section 2). But in practice, there are more efficient methods by which this result may be accomplished. To these we now turn.

Combinations. Any group or set of distinct objects is called a combination, symbolized C. Such a joint event may consist of a list of digits, a handful of coins, a group of nine baseball players, a series of births, a set of postage stamps, a bridge hand, a list of True–False questions, or a sample of a human population. The substitution, addition, or subtraction of even a single item in the given set produces a different combination. Thus, the two sets of digits, 2698 and 2697, constitute two different combinations. The items must be distinct, but need not be distinguishable, as in 4444 and 44444, which are different combinations. For a given problem, a set of five pennies may be considered as indistinguishable, although each bears a different date — a fact which may be, however, important to a numismatist.

Permutations. The above illustrations make it evident that in many situations we are interested not only in the identity of the combinatorial elements, but also in the order of their arrangement. A telephone number of 2698 is not identical with 2986; the same baseball nine may be put forth in different batting orders; five classmates may be chosen in different orders of sociometric preference; the pins in a Yale lock may be placed in different orders so that only the corresponding key will fit it. In the above cases the number and identity of the items remain unaltered; they remain the same *combination*, but they take on different *permutations* because two or more items in the arrangement are re-ordered. When the focus of attention is on the order of things, the series is a permutation, symbolized by P.

Popular language does not always correctly distinguish these two terms. In order to open a bank vault lock, consisting of gears and pins which must be actuated in a specific sequence, the operator must know, strictly speaking, not merely the combination, but also the permutation of the dial numbers. A rearrangement of any two (or more) of the pins would produce another permutation and make it impossible to open it.

The Calculation of Permutations. Any or all of the objects in a combination may be permuted, provided the objects are distinguishable. This assuming that all marry and that religon of husband and wife are independent.

	Female	*Male*
Protestant	20	30
Jewish	60	50
	80	80

5. Given a group of 100 persons, of whom 20 are Republicans and 80 Democrats. If 5 are selected by chance, what would be the probability that all would be Republican?

6. In a large number of throws of two true dice, how often would you expect a total of seven?

7. If two dice were repeatedly thrown, which event would surprise you more: a double-six on the first throw, or on any one of the subsequent throws?

8. Comment on the statement: "All individuals in the group are by definition subject to the same probability statement; hence, the individual, in the cohort for which the prediction is made, cannot justifiably exempt himself from the operation of the quoted probability."

9. If the Income Tax (Internal Revenue) Office selects a 25 per cent random sample of the returns for checking, what would be the probability of a person being selected for checking two years in succession? Comment.

10. You have bought 10 chances, at $.25 each, on an automobile which is being raffled for charity. Suppose you wished to sell 5 of these raffle tickets. What would be a fair price for a person not interested in charity? Discuss.

SECTION THREE

Combinations and Permutations

A joint event is, by definition, a grouping of individual events. It could comprise the letters in a word, a bridge foursome, or a group of musical notes. But groups can usually be assembled in more than one way. Thus, there are 36 ways in which two dice may fall; a 3 and a 5 may occur in two ways since each figure can appear on either one of two dice, as has been previously shown. Similarly, letters can be scrabbled — arm, ram, and mar; four persons may be seated in various orders around a table; the same musical tones may take various positions in a chord or in a melodic sequence. A facility in such regrouping is of prime importance in the calculation of probabilities, because the magnitude of a joint probability is necessarily increased in proportion to the number of ways in which a specified joint event can occur.

ways. If all ten digits (0–9) were available, the computation would be $10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 3{,}628{,}800$. The principle here involved is referred to as the *Multiplication Theorem*, the generalized formulation of which is:

> If an event can occur in N_1 ways, and thereafter in N_2 ways, and so on, these successive events can occur in that order in $N_1 \times N_2 \times \cdots N_N$ ways.

This theorem is employed in both of the fundamental patterns of permutation: (1) complete, and (2) partial.

Complete Permutation, or Factorial. A complete permutation consists in permuting all objects in a set, all at a time: that is, N objects, N at a time. The number of permutations resulting therefrom is called the *factorial of the number*, and is symbolized:

$$P_N^N = N! \tag{8.3.1}$$

In permuting four distinguishable objects, $4!$ would cumulate to 24 different orders, as has already been illustrated. A factorial of a number (N) is therefore defined as the product of all integers N to 1, inclusive.

Partial Permutation. A permutation of N distinguishable objects, less than N at a time, may be called partial permutation, symbolized by P_r^N, to be read "the number of permutations of N objects, r at a time." This formula is a directive to carry out the permutations only through r stages, instead of through the total N.

$$P_r^N = N(N-1)(N-2)\cdots(N-r+1) \tag{8.3.2}$$

For example, the number of four-digit telephone numbers, when all ten digits are available (but when the same digit is allowed to appear only once in any given telephone number), would be derived as follows:

$$P_4^{10} = 10 \times 9 \times 8 \times 7 = 5{,}040$$

Recurring Items. Combinations which contain indistinguishable items are amenable to permutation only in proportion to the number of distinguishable objects. Nevertheless, there are certain combinations in which the recurrence of items is not only permitted, but even required, as, for instance, in the case of telephone numbers and computer punch cards.

In the case of discrete physical objects, such recurrences are obviously impossible. In permuting nine boys on a baseball field, physical circumstance precludes their recurrence in successive positions on the same "trial." A boy cannot play first base and center field in the same lineup. The same principle holds for sociometric arrangements. However, numerical symbols and letters of the alphabet do not suffer from this limitation; hence, if one wishes to determine the total number of all possible proviso is quite logical since a re-ordering of indistinguishable objects is meaningless: a rearrangement cannot be recognized insofar as the objects are considered to be absolutely identical.

If the number of objects in a combination is small enough, the permutations can be formed by inspection and counted. Thus, the three digits, 2, 5, and 6, can be permuted in six ways as follows:

256	562	625
265	526	652

The addition of one digit to the set (2568) would not change the number of potential combinations, but would increase the number of possible permutations from six to 24, as the *tree diagram* in Figure 8.3.1 will demonstrate.

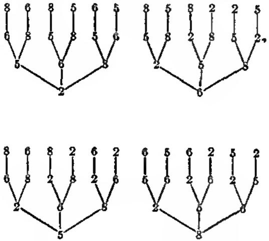


FIGURE 8.3.1 *Tree Diagram of Permutations*

But the same result could have been more quickly achieved without such extensive and detailed figuration by simple multiplication according to the following logic:

Examination of the foregoing 24 permutations reveals that the first digit in a permutation could be any one of the four — that is, the first digit could occur in four ways. After each of these four possibilities, the next digit could fall in any one of three remaining ways to join each of the preceding four, yielding so far twelve different permutations, two digits at a time. Thus, the pattern of calculation could continue until all the available digits are used up as above. Summarizing the arithmetic procedure to count the number of ways, we have: $4 \times 3 \times 2 \times 1 = 24$

According to this formula, the number of different ways in which we may select 4 items from a pool of 10 would be:

$$C_4^{10} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1}$$
$$= 210 \text{ combinations}$$

These combinatorial methods are of special significance in the calculation of the binomial probabilities, which is the subject of the following section.

Questions and Problems

1. Define the following concepts:
   Combination
   Permutation
   Multiplication Theorem
   Recurring Items
   Tree Diagram

2. A population can be classified by race in 3 ways, by religion in 3 ways, and by nativity in 2 ways. Prepare a tree diagram to show in how many ways persons may be classified simultaneously.

3. How many telephone numbers of 4 digits each can be made from all digits 0 to 9 if: (a) no telephone number begins with zero? (b) no duplicate digits are permitted? (c) all digits are permitted to recur?

4. A recreation director wants each boy on a baseball team to have the opportunity to play each position for a full game.
   (a) How many games must he play?
   (b) How many games would be have to play if he used all possible batting orders?

5. In the columns of an IBM card, each labeled 0 to 9 inclusive, how many different code numbers can be made using 1, 2, or 3 columns?

6. It is usually assumed that additional culture traits (discoveries or inventions) increase the possibilities of new inventions, since new traits may combine with any one of a large number of already existing traits. In principle, this is a statistical approach to the problem of culture growth. Assuming that all traits may freely combine, how many permutations could be formed with 2, 3, 4, 5, or 6 traits? *What generalization does this suggest?*

7. What would be the more probable bridge hand. any specified set of 13 cards, or a complete suit? Explain.

8. In a toss of ten coins, why are ten heads (tails) so much less probable than any one of the other combinations? (*Hint:* Calculate C_r^N for $N = 10$, and all possible values of r.)

arrangements, with no restriction on recurrences, one must allow the multiplication theorem to take over from the factorial principle. Whereas all factorials are based on the multiplication theorem, not all cases in which this theorem is employed need be factorials, as in the instance of telephone numbers. The number of possible four-digit telephone numbers, allowing all ten digits to recur, would be $10 \times 10 \times 10 \times 10 = 10{,}000$. The symbolism for such an operation is N^r, where r is the number of positions to be filled, in this instance 10^4. If the N objects are to be serialized in N positions, the symbol would be N^N.

Calculation of Combinations, r at a Time. There is no difficulty in the calculation of the number of combinations, when N objects are taken N at a time, for there can be only one combination. A problem ensues only when fewer than N objects are used to form the combinations — when we take them r at a time.

As in the case of permutations, when N is small, the subsidiary combinations can be found by inspection. A group of four children, A, B, C, D, taken two at a time, could combine in the following six ways:

AB	BC	CD
AC	BD	
AD		

The partial permutations of the same group, two at a time, would total to 12. Clearly, there will always be fewer combinations than permutations, for any given N and r.

With larger N's, it is tedious to determine the number of combinations by means of a tree diagram. Instead it is more convenient to compute first the number of permutations of N objects, taken r at a time. The second step consists in clearing the result of these permutations by dividing by $r!$, which is the number of ways each combination can be permuted. In other words, since each combination can have $r!$ permutations, we may obtain the number of combinations by dividing the number of permutations by $r!$ This division obviously reduces the permutations to the corresponding number of combinations. The formula is as follows:*

$$C_r^N = \frac{P_r^N}{r!} \tag{8.3.3}$$

where N = total number of items
r = number of items in combination

* In many texts in college algebra, as well as in statistics, the formula is written:

$$C_r^N = \frac{N!}{r!\,(N-r)!}$$

This formulation is obtained by multiplying both members of the fraction by $(N-r)!$, thus producing complete permutations in both the numerator and denominator, which is considered by some to be a simpler procedure.

$$\begin{aligned} pq &= \frac{1}{2} \times \frac{1}{2} \\ &= \frac{1}{4} \end{aligned}$$

where p = the probability of a male
q = the probability of a female

But the outcomes may occur in reverse order — the girl on the first trial and the boy on the second trial — making a total of two ways in which the specified joint event can occur. Therefore, the probability of a male birth and a female birth, irrespective of the order of outcomes would be

$$\begin{aligned} Pr(m, f) &= \frac{1}{4} + \frac{1}{4} \\ &= 2(\tfrac{1}{4}) \\ &= \frac{1}{2} \end{aligned}$$

Thus, the probability of a specified joint event is the probability of any given joint event weighted by the number of ways in which it can occur.

Similarly, we may establish the probability of guessing 3 right and 2 wrong on a True–False test of five items. We first find the probability of 3 right and 2 wrong in that order:

$$\begin{aligned} pppqq &= p^3q^2 \\ &= (\tfrac{1}{2})^3(\tfrac{1}{2})^2 \\ &= \tfrac{1}{32} \end{aligned}$$

where p = probability of guessing right
q = probability of guessing wrong

Next, we determine the number of ways in which three rights (R) and two wrongs (W) could occur on five trials. By listing, we discover a total of 10 ways:

RRRWW
RRWRW
RWRRW
WRRRW
WRRWR
WRWRR
RRWWR
RWWRR
RWRWR
WWRRR

Combining the two results, we arrive at a probability of $\tfrac{10}{32}$ *for the specified* joint outcome, 3 rights and 2 wrongs.

Full Set of Binomial Probabilities. By identical logic, we could obtain the probabilities of 2 and 4 rights, respectively. The probability of all 5 right, hy guessing, would be $\tfrac{1}{32}$, since there is only one way of obtaining 5 successes on five trials. Likewise, the probability of all wrong on five

SECTION FOUR

Binomial Probabilities: Independent Events

The Binomial. Every possibility set obviously may be reduced to two mutually exclusive possibilities, and the resulting outcomes labeled *success* and *failure.* The outcome termed success may be an elementary event, such as a male birth, or it may be one of a group of elementary events, such as four aces in a deck of playing cards. In tbese instances, the probabilities of success would be $\frac{1}{2}$ and $\frac{4}{52}$, respectively. But however defined, whenever the possibility set consists of only two alternatives, there can be correspondingly only two probabilities: the probability of success, p, and the probability of failure, q. Hence, tbe probabilities of the twofold set may always be written:

$$(p + q) = 1.00$$

Sincs ths expression on the left is the algebraic sum of two terms, it is a *binomial.**

Now, while ths layman's interest is almost always centered on ths probability of success on a single trial, the statistician's interest usually lies in the *probability of r successes on N trials. He is interested in the* general quantitativs laws that govern tbe behavior of events on repeated trials, and which enable him to predict the likelihood of their occurrencs. For example, he may desirs the probability of two heads on successive throws, or the probability of a 5 and 5 when two dice are thrown together. We have already shown that the probability of a 5 and 5 is $\frac{1}{36}$, since a 5 and 5 may occur in only one way. On the other band, the probability of a 3 and a 5 is $\frac{2}{36}$, since a 3 and 5 may occur in two ways. Analogously, the probability that the faces of two dice will sum to 7 is $\frac{6}{36}$, and the probability that they will sum to 8 is $\frac{5}{36}$.

Clearly, a specified outcome may occur in more than one way. Hence, in order to calculate the probability of that outcome, we must always determine the number of ways in which it can occur; otherwise we cannot properly weight its probability. As we have done before, we pursue a very simple example: the probability of parents having a boy and a girl (ignoring multiple births). We first calculate the probability of a male birth and a female birth *in that order.* This is, according to the product theorem for independent events:

* In the analysis that follows, outcomes are assumed to be independent from trial to trial — that is, the probabilities of success and failure are constant from trial to trial

This probability series naturally sums to unity, since every possible joint outcome is accommodated berein.

Binomial Probability Distribution. Since probability is identical with relative frequency, tbe binomial probabilities may be viewed as the frequency distribution of the variable "r successes on N trials," r taking all integral values from zero througb N. Such a distribution is appropriately entitled a *binomial probability distribution,* and may be displayed in either tabular or graphic form. Tbus; where $p = \frac{1}{2}$, $q = \frac{1}{2}$, and $N = 5$, as in the foregoing example, the binomial probability table would be constituted as shown in Table 8.4.2. The corresponding bistogram is shown in Figure 8.4.1.

Table 8.4.2

Binomial Frequency Table, $N = 5$, $p = .5$

r	$Pr(r)$
5	$\frac{1}{32}$
4	$\frac{5}{32}$
3	$\frac{10}{32}$
2	$\frac{10}{32}$
1	$\frac{5}{32}$
0	$\frac{1}{32}$
	1.00

Mean and SD of Binomial Distribution. By definition, the mean of a binomial distribution is the average number of successes on all trials; it is therefore the expected number per trial. The *SD* is, of course, an average of chance variation around that mean expectation. Both quan-

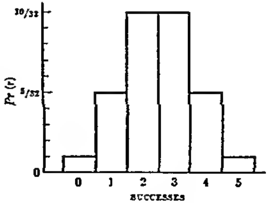


Figure 8.4.1 *Histogram, Binomial Probability Distribution ($N = 5$, $p = .5$)*

trials would be $\frac{1}{32}$, there being only one possible way of getting every item wrong. All of these results may be usefully assembled in a single display which suggests the law of their formation (Table 8.4.1).

Table 8.4.1 *Combinations of Rights and Wrongs, Five True-False Items*

	NUMBER RIGHT						TOTAL
	5	4	3	2	1	0	
	RRRRR	RRRRW RRRWR RRWRR RWRRR WRRRR	RRRWW RRWWR RWWRR WWRRR WRWRR WRRWR WRRRW RWRWR RWRRW RRWRW	WWWRR WWRRW WRRWW RRWWW RWRWW RWWRW RWWWR WRWRW WRWWR WWRWR	WWWWR WWWRW WWRWW WRWWW RWWWW	WWWWW	
Combinations	1	5	10	10	5	1	32
Pr(R)	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$	1.00

From an inspection of this complete set of binomial probabilities, we may discern the general rule: to find the probability of r successes in N independent trials, find the probability of r successes and $N - r$ failures in that order, and then weight that probability by the number of ways in which r successes can occur. But, according to the rule on combinations, the drawing of r successes from N trials is simply C_r^N Hence, the probability of r successes on N trials is:

$$Pr(r \text{ successes}) = C_r^N p^r q^{N-r} \tag{8.4.1}$$

The Expanded Binomial. Were we to express each probability in the above notation and arrange them in order from five successes to none at all, we would have the following sequence:

$$C_5^5p^5q^0,\ C_4^5p^4q^1,\ C_3^5p^3q^2,\ C_2^5p^2q^3,\ C_1^5p^1q^4,\ C_0^5p^0q^5$$

And, were we to sum these terms, we would have an instance of the *expanded binomial*, since the algebraic results correspond to raising the binomial $(p + q)$ to the Nth power, $(p + q)^N$; in our example, $(\frac{1}{2} + \frac{1}{2})^5$:

$$\tfrac{1}{32} + \tfrac{5}{32} + \tfrac{10}{32} + \tfrac{10}{32} + \tfrac{5}{32} + \tfrac{1}{32} = 1$$

Table 8.4.4

Binomial Probabilities, $N = 10$, $p = .05$

Number Failing (r)	$Pr(r)$
10	$1(.05)^{10}(.95)^{0}$
9	$10(.05)^{9}(.95)^{1}$
8	$45(.05)^{8}(.95)^{2}$
7	$120(.05)^{7}(.95)^{3}$
6	$210(.05)^{6}(.95)^{4}$
5	$252(.05)^{5}(.95)^{5}$
4	$210(.05)^{4}(.95)^{6}$
3	$120(.05)^{3}(.95)^{7}$
2	$45(.05)^{2}(.95)^{8}$
1	$10(.05)^{1}(.95)^{9}$
0	$1(.05)^{0}(.95)^{10}$

flunk on each trial is 5 per cent — or, in other words, the probability of any given student failing is 5 per cent. Nevertheless, in an extreme case, the whole class of 10 may conceivably flunk — but not very often — and still be consistent with the theoretical average of 5 per cent failures. The probability that the whole class would fail is represented by the first term of the expansion: $(.05)^{10} = \frac{1}{10{,}240{,}000{,}000{,}000}$. Obviously, this is an extremely improbable event. However, a freak event does occasionally occur, however miraculous it may seem when it happens to you. This particular event is only 15 times as unlikely as a specific combination of cards in any given hand of bridge.

Binomial and Normal Distributions. By means of the binomial expansion it is a simple matter to determine, for example, the probabilities of 3, 2, 1, and 0 boys in families of three children. However, this method of calculating probabilities becomes cumbersome when the number of trials is large — for example, when we wish to calculate the probability of 90 or more beads in 100 throws or of 60 or more rights by guessing on a True–False test of 100 items. Confronted by such problems, early statisticians posed the question of whether the binomial distribution approaches some fixed pattern as N gets larger, which in turn might be used to provide approximately the desired probabilities.

This interesting and fruitful possibility was explored and resolved by De Moivre over 200 years ago in his discovery that the binomial distribution more and more nearly resembles the normal curve as N increases without limit. *In fact, N has only to exceed 20 in order to produce a* very good fit between binomial and normal distributions, provided that p is not too divergent from $\frac{1}{2}$. This convergence of the discrete binomial tities may be directly computed in the usual manner. Thus, to find the mean, we (1) weight each value by its corresponding probability (frequency); (2) sum these weighted values; and (3) divide that sum by the total probability (frequency). Since the sum of the probabilities is always unity, the mean of the binomial probability distribution is simply the sum of the values weighted by their respective probabilities. The *SD* is analogously computed. In Table 8.4.3, we illustrate the computation of the mean of the binomial frequency distribution of Table 8.4.2.

Table 8.4.3

Computation of Mean, Binomial Frequency Distribution

r	$Pr(r)$	$Pr(r) \times r$
5	$\frac{5}{32}$	$\frac{5}{32}$
4	$\frac{5}{32}$	$\frac{20}{32}$
3	$\frac{10}{32}$	$\frac{30}{32}$
2	$\frac{10}{32}$	$\frac{20}{32}$
1	$\frac{5}{32}$	$\frac{5}{32}$
0	$\frac{1}{32}$	0
	1.00	$\frac{80}{32} = 2.5$
$\bar{X} = \frac{2.5}{1.00} = 2.5$		

Practically, such cumbersome calculations are unnecessary, since the mean of the binomial distribution of r successes on N trials is always p of N, usually written Np, and the *SD* is $\sqrt{pq \text{ of } N}$, usually written $\sqrt{Npq}$. Applying these formulas to the above data we obtain:

$$\bar{X} = Np \qquad (8.4.2)$$
$$\bar{X} = 5(\tfrac{1}{2}) = 2.5$$

$$\sigma = \sqrt{Npq} \qquad (8.4.3)$$
$$\sigma = \sqrt{5(\tfrac{1}{2})(\tfrac{1}{2})} = \sqrt{1.25} = 1.1$$

Empirical Probabilities. The binomial expansion need not result in a symmetrical distribution, nor need it refer only to *a priori* probabilities. If, for instance, over a period of time, 5 per cent of the freshmen in English I fail to pass, 95 per cent would succeed in passing. Presumably, future students would conform to those probabilities. The corresponding binomial would therefore be: $(.05 + .95) = 1$. Accordingly, we may ask illustratively with what frequency the various combinations of successes and failures would occur by chance in the long run in classes of 10 students. The solution is provided by the successive terms of the expanded binomial shown in Table 8.4.4. In this situation, a successful

Areas. We need only to express an observed *r*-value as a standard deviate, and then to evaluate that standard deviate by means of this table. Thus, to find the probability of guessing 60 items *or more* right on a True–False test of 100 items, we first express 60 as a standard deviate and then determine the frequency with which more exceptional values are expected to occur. To convert 60 to standard form, we require of course the mean and *SD* of the distribution:

$$\begin{aligned} \bar{X} &= Np \\ &= 100(\tfrac{1}{2}) \\ &= 50 \end{aligned} \qquad \begin{aligned} \sigma &= \sqrt{Npq} \\ &= \sqrt{100(\tfrac{1}{2})(\tfrac{1}{2})} \\ &= \sqrt{25} \\ &= 5 \end{aligned}$$

By means of these values, we convert 60 to a standard deviate in the usual manner, with one minor adjustment. Since the normal curve is based on a continuous variable, we treat discrete 60 as the midpoint of an interval extending from 59.5 to 60.5. Since our problem is to determine the probability of 60 or more right by guessing, rather than more than 60, our standard deviate is calculated on 59.5 instead of 60.5.

$$\begin{aligned} \frac{r - Np}{\sqrt{Npq}} &= \frac{59.5 - 50}{5} \\ &= 1.9 \end{aligned}$$

From the table we find that approximately 3 per cent of all items in a normal distribution lie beyond 1.9 sigmas; hence, the probability of obtaining 60 or more by guessing is approximately 3 per cent. In other words, in a class of guessers, 3 per cent could expect to score 60 or higher.

Although 3 per cent of a large class of students could be expected to answer correctly 60 or more out of 100 True–False questions by sheer chance, we cannot infer from such an observed score that they actually did guess. An obtained score of 60, for example, could simply mean that the student controlled 60 per cent of the subject matter. How, then, is the professor, or any other observer, to determine the correct explanation of such a score? The answer is: on the basis of the statistical evidence, he simply cannot know for sure. The central issue of what is known in statistical interpretation as *decision-making* is now joined: whether to accept the hypothesis of chance (also known as the *null hypothesis*), or to explain the score of 60 by the presence of some determining factor such as studiously acquired knowledge or possibly even cheating. There is no formula for adjudicating between these two alternatives, although as the probability of the chance event becomes smaller and smaller, the chance hypothesis of guessing becomes correspondingly less and less acceptable. Such questions will be treated in Chapter 12 on statistical inference.

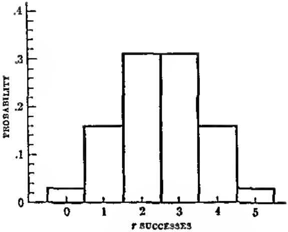


FIGURE 8.4.2a *Histogram, Probability Distribution* $(N = 5, p = .5)$

on the continuous normal distribution is illustrated in Figures *8.4.2*a and 8.4.2b, where $p = \frac{1}{2}$, and $N = 5$ and 10, respectively.

The practical significance of this equivalence between the binomial and normal probabilities inheres in the fact that the binomial probabilities may be handily obtained from the usually accessible Table of Normal

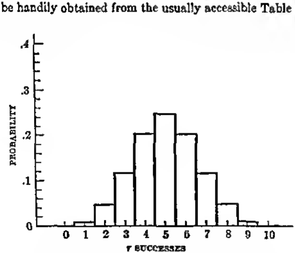


FIGURE 8.4.2b *Histogram, Probability Distribution* $(N = 10, p = .5)$

§ 8.4 Binomial Probabilities: Independent Events

Feller, William, *An Introduction to Probability Theory and Its Applications*, 2d edition. John Wiley & Sons, Inc., New York, 1957. Volume I, Chapters 1 and 2.

Kemeny, John G., J. Laurie Snell, and Gerald L. Thompson, *Introduction to Finite Mathematics*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1956. Chapters 3 and 4.

Laplace, Marquis Pierre Simon de, *A Philosophical Essay on Probabilities*. Dover Publications, Inc., New York, 1951.

Levinson, Horace C., *Science of Chance*. Rinehart & Company, Inc., New York, 1950. Chapters 14 and 15.

Nagel, Ernest, "Principles of the Theory of Probability," Vol. 1, No. 6, *International Encyclopedia of Unified Sciences*. The University of Chicago Press, Chicago, 1939.

Neyman, Jerzy, *First Course in Probability and Statistics*. Henry Holt and Company, New York, 1950. Chapters 1 and 2.

Polya, George, *Patterns of Plausible Inference*, Vol. II of *Mathematics and Plausible Reasoning*. Princeton University Press, Princeton, N.J., 1954. Chapters 14 and 15.

Savage, Leonard J., *The Foundation of Statistics*. John Wiley & Sons, Inc., New York, 1954. Chapter 4.

Venn, John, *The Logic of Chance*. The Macmillan Company, London and New York, 1876.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Binomial
   Binomial Probabilities
   Expanded Binomial
   Binomial Probability Distribution
   Independent Outcome

2. It has been contended that the excess of males over females at birth in the United States is due to the tendency of families to have no children after a male birth. Would such a parental bias affect the sex ratio?

3. A person claims he can discriminate between the tastes of two cigarettes. He is given Brands A and B in pairs, 5 times. Assuming he merely guesses, calculate the probability of 5 correct choices; 4 correct choices; at least 4 correct choices. What chance factors would interfere with his judgment? How many rights would you require to be convinced that he has the power to discriminate?

4. Are the following probabilities identical: 5 heads in 5 tosses of 1 coin; 5 heads in 1 toss of 5 coins? State the rule.

5. In tossing a coin, how long a run of heads could one throw without arousing suspicion of a biased coin? Explain your reasoning.

6. In a throw of two coins, what is the probability of obtaining exactly 50 per cent heads and 50 per cent tails? Calculate also for 4, 6, 8, and 10 coins. What happens to the probabilities of an even split as the number of coins is increased from two to ten?

7. If a True–False examination consists of six questions, what is the probability that the student will mark by chance:
   (a) all correctly?
   (b) half of them correctly?
   (c) all incorrectly?

8. Does the expanded binomial always result in a symmetrical distribution? Calculate the distribution of fives and no-fives with 2, 4, and 5 dice. What generalization can you offer? (*Hint:* Prepare a histogram.)

9. The Old Testament states that Jacob had 12 sons. We may hypothesize either (a) that there were actually no daughters among the children, or (b) that the daughters were not recorded. Determine the probability of the first alternative, and state which assumption you favor. What other factors, besides statistical evidence, might enter into your decision?

SELECTED REFERENCES

Boole, George, *An Investigation of the Laws of Thought*, unabridged reproduction of the 1854 edition. Dover Publications, Inc., New York. Chapter 16.

Although no event is conceivable in isolation, it is not a simple matter to detect with what variables it is linked. For example, the association between delinquency and intelligence was once thought to be very close, but in later analysis was found to be quite remote. This shift in interpretation was a result not only of more carefully controlled studies of delinquency and its associated social factors, but also of a redefinition of the test-intelligence norms themselves.

To man, nature is infinitely complex; hence, the search for such underlying relations can never be completed. There are innumerable factors whose relations are so intricate that the "final" factor in the production of an event can never be attained, much less measured. However, in its unfulfilled quest for certainty, our common sense begins very early to construct patterns of relationships on the basis of which it endeavors to comprehend the past, understand the present, and thereby anticipate the future.

Patterns of Relationship. The patterns of association are, of course, continuously revised in the trial and error of daily experience. In the process, the observer mentally quantifies and summarizes his observations by first noting the factors which seem to him to "cause" or to be linked with the event and, second, by noting the frequency with which he successfully anticipates or forecasts it. In this casual manner, every person begins to cultivate the habits of association and becomes an informal statistician. He practices intuitively the principle of correlation. In fact, much sociology effectively employs such intuitive statistics, skillfully put together by alert, widely-traveled scholars, made without benefit of pencil and paper calculation or technical procedures. Indeed, many who are critical of the utility of statistical procedures nevertheless unwittingly employ them in this unofficial manner.

For some purposes, a rough subjective approximation of a correlation is fairly satisfactory, but for scientific purposes more accurate measurements are desired. Such precision is not a simple matter to achieve. The difficulty lies in the complexity of the patterns of relationship in terms of which we view and organize the world. Some of the salient features of this complexity in patterns may be formulated as follows: (1) every event is the outcome of multiple factors; (2) the force of these respective factors varies in intensity, and (3) may flow in one or more directions in producing its effects; (4) the factors are in constant interaction, and consequently (5) they may reinforce, counteract, and cancel one another. We illustrate with a popular example: the halfback in a football game, whose scoring power tends to be associated positively with running ability, is either obstructed or aided in actual performance by the condition of the field, his fatigue, the type of plays called, to say nothing of the varied activities of the other twenty-one players on the

Measurement of Association: Qualitative Variables

9

Section One

Concept of Statistical Association

Principle of Contingency. It is an axiom of science, as well as of common sense, that no event in nature "just happens," but always occurs under very specific, known or unknown, circumstances. An event is therefore never to be viewed in isolation. It must be considered as a product of the joint operation of many forces, each of which contributes a variable element to the observed outcome. Thus, the size of family may be dependent on such factors as age at marriage, level of income, extent of employment of the mother outside the home, and the religious ideology entertained by the parents. Parole success of released prisoners may be related to the type of crime committed, age, and history of recidivism. Some of these factors tend to promote and accelerate, others to retard in varying degrees or even inhibit, the occurrence of the event, or at least to modify its character or magnitude. Thus, the religious factor may foster large families, while the economic factor may restrain that tendency. In any case, we cannot predict the degree of parole success, or explain the size of the family, unconditionally, but only on the basis of specifically designated factors, or variables, on which the outcome is contingent. The human observer does not possess absolute knowledge; he must use one event as a cue to anticipate another. His understanding is therefore grounded in the *principle of contingency*. And it is in accordance with that principle that the sciences, both physical and social, have set up their methods and objectives: (1) to identify the variables (determining factors) associated with an event; (2) to uncover the patterns of association between the variables and the event; and (3) to measure the strength of that association.

field. This association between scoring power and running ability may therefore not be visible to the fans in the stadium. Just so is the correlation between income and size of family beclouded by such factors as religion, age at marriage, occupation of breadwinner, and education. It may at first seem a hopeless task to disentangle these networks of relationships and to subject them to statistical reasoning. Nevertheless, in a large number of observations, the essential relation can be expected to shine through the haze formed by tha multiplicity of factors.

This problem is probably less difficult in the physical than in the social sciences. The physical scientist, by means of available laboratory controls, is to a certain extent able to segregate and manipulate his elements and to replicate his careful and undisturbed observations, whereas the social scientist is often obliged to accept data which are like unrefined ores from "nature in the raw" and to assemble materials from widely dispersed sources and a variety of settings. He is therefore compelled to employ statistical controls, since, in general, laboratory controls are closed to him.

What Are the Evidences of Relationship? How can we be sure that factors are interconnected? And, having discovered a relationship, how may we determine the degree or intensity of that relation? Broadly speaking, there are two earmarks of such linkage: (1) joint occurrence of attributes, and (2) parallel changes in two or more series of quantitative observations.

The relative frequency with which certain attributes happen together is probably the most elementary basis of lay judgment of association. This is the *principle of joint occurrence.* Statistical variables, like human beings, are usually judged "by the company they keep." For example, if delinquency is more often found in boys than in girls, we conclude that delinquency is associated with "boyness." The strength of this association will vary according to other factors such as the boy's age, the type of delinquency, and many other elements in the pattern, all of which will render the statistical application of an apparently simple principle quite complicated. Hence, it need hardly be reiterated here that some system of tabulation and classification is necessary as an aid not only in establishing an association, but also in determining its strength.

Second, if in two series of quantitative data, a unit change in one variable is paralleled with some degree of regularity by a comparable change in the other series — that is, if they move together — we conclude that they are somehow tied together, and that there is an association between the two sets of data. For example, as income declines, the size of family tends to increase; and, if the observations endure through a rather extensive range — that is, for the entire range of families of all sizes and of incomes of varying amounts — the evidence of a relation is strengthened. This is called the *principle of covariation.*

But the important point is that delinquencies are distributed between boys and girls, and that auto occidents are shared by both men and women drivers. We must therefore inquire in what proportion the limited supply of violations or auto accidents are divided between the sexes. To which sex do delinquencies or accidents preferentially attach themselves? To estimate the degree of affinity for *either* sex, we must know the delinquency rate for both sexes; a delinquency rate for boys can be said to be high only when the rate for the girls is known. Let us therefore provide the rates for females and Democrats in the 2 × 1 tables, and analyze the results.

Table 9.2.2 *2 × 2 Tables*

	Boys	*Girls*		*Rep.*	*Dem.*
Delinquent	10%	0%	Isolationist	50%	30%
Non-delinquent	90	100	Internationalist	50	70
	100%	100%		100%	100%
	Male	*Female*		*Men*	*Women*
A Grades	10%	20%	Accident	10%	20%
Other	90	80	No Accident	90	80
	100%	100%		100%	100%

Source: Hypothetical

The delinquency rate of boys now turns out to be rather high, since even a 10 per cent delinquency rate is higher than no delinquency at all. Similarly, the grades of college men are below the norm; the Republicans display a marked propensity toward isolationism; and women drivers show a marked susceptibility to mishaps.

The fact is that we have now introduced a standard of judgment against which the 2 × 1 table may be compared. Some standard is inevitably and unwittingly introduced by every observer, and it is the function of statistical procedures to make the standard explicit. Hence, a 2 × 2 table is the minimum for a dependable conclusion on association.

The 2 × 2 Table. A 2 × 1 table merely presents the two subclasses of a single variable, whereas a 2 × 2 table presents the subclasses of two variables cross-classified by one another. In Table 9.2.3, a case is classified as boy and delinquent, as girl and non-delinquent. This double classification automatically establishes the distribution of joint occurrences between a given sex and behavior which is, of course, basic to the understanding of the phenomenon of association. In the hypo-

Section Two

Coefficient of Association: Yule's Q

One of the simplest measures of association is known as the *Coefficient of Association*, or, more informally, as *Yule's Q*. This measure is designed to reflect the degree of association between a pair of qualitative variables, arranged in a 2 × 2 (fourfold) table. In his study of cross-classification the student has already acquired the sense of inferring association, and even cause-and-effect relationships, between dichotomous variables. But at that juncture, we did not seek to compute a single over-all measure which would reflect the strength of that relation. Yule's Q and other measures of association are designed to do just that, since they are summarizing measures for biveriate data, enalogous to the mean for univariete data.

Inadequacy of the 2 × 1 Table for Estimate of Association. Many persons untrained in quantitative reasoning naively succumb to the temptation of drawing conclusions on the degree of association from a 2 × 1 table, instead of from a 2 × 2 table, which is the minimum for that purpose. The deceptive ease with which such erroneous deductions can be made is illustrated in the four 2 × 1 tables depicted in Teble 9.2.1.

Table 9.2.1 *2 × 1 Tables*

Boys	*Per Cent*	*Republicans*	*Per Cent*
Delinquent	10%	Isolationist	50%
Non-delinquent	90	Internationalist	50
	100%		100%
Males	*Per Cent*	*Female Drivers*	*Per Cent*
A Grades	10%	Accident	20%
Other	90	No Accident	80
	100%		100%

Source: Hypothetical

Since only 10 boys out of 100 are delinquent, a layman may be misled into the conclusion that there is a very weak association between boyness and delinquency; similarly, that women are not accident prone. Since Republicans divide 50-50 on isolation, there would seem to be no striking tendency for Republicans to favor one or the other foreign policy.

sies in reflecting gradations of association. Yule's Q, being our first example of a measure of association, in its simplicity supplies us with a convenient introduction to this type of analysis which will, in its turn, throw considerable light on all subsequent instances. How sensitive, then, is Q to the variations in the data?

First let us assume that not all the delinquents are boys, but rather that 1 and 5 girls in 50, respectively, are delinquents. (See *a* and *b* of Table 9.2.4.) In both instances, there is still a positive association between delinquency and boyness, but not as complete as in the first table. Thus, the value of the index is reduced; it reaches zero when a delinquent is just as likely to be a boy as a girl. Sex and delinquency are then said to be independent; pure chance prevails. Statistically speaking, the internal cell ratios are identical with the corresponding marginal ratios. You might as well toss a coin for your prediction; hence, $Q = 0$.

Table 9.2.4 *Sensitivity of Q*

	(*a*)			(*b*)			(*c*)		
	B	G		B	G		B	G	
Del.	19	1	20	15	5	20	10	10	20
Non-del.	31	49	80	35	45	80	40	40	80
	50	50	100	50	50	100	50	50	100
	$Q = .91$			$Q = .59$			$Q = 0$		

Effect of Marginal Ratios. We have seen how Q reflects the changes in row (column) ratios, with marginals remaining constant. However, Q is not at all sensitive to changes in marginal ratios, so long as the corresponding cell ratio remains fixed. For purposes in hand, let us construe the column subtotals as samples of 100 boys and 50 girls (Table 9.2.5).

Table 9.2.5 *Stability of Q*

	(*a*)			(*b*)			(*c*)		
	B	G		B	G		B	G	
Del.	10	3	13	10	6	16	20	6	26
Non-del.	90	47	137	90	94	184	90	47	137
	100	50	150	100	100	200	110	53	163
	$Q = .27$			$Q = .27$			$Q = .27$		

Table 9.2.3

Number of Delinquents by Sex

	Boy	Girl	*Total*
Delinquent	20	0	20
Non-delinquent	30	50	80
Total	50	50	100

thetical group of 100 children shown in Table 9.2.3, there are 20 delinquents, all of whom are boys. Most of the children are, of course, nondelinquent. But, given a delinquency, there is perfect prediction — that *is*, complete certainty about the sex of the delinquent. *Hence*, taking the data at face value, we may say that delinquency is completely explained by boyness, since there is obviously no element of girlness required for its occurrence. The completeness of this explanation should register itself in any index that may be contrived for that purpose.

In order to provide a single measure of association for such a 2 × 2 table, the English statistician, G. Udney Yule, proposed the following coefficient of association, which he labeled "Q" in honor of the nineteenth century statistician, Quételet:

$$Q = \frac{ad - bc}{ad + bc} \tag{9.2.1}$$

where a, b, c, d are the joint frequencies conventionally arranged in the fourfold table in the following manner:

a	*b*
c	*d*

Applying the formula to Table 9.2.3:

$$\begin{aligned} Q &= \frac{(20 \times 50) - (30 \times 0)}{(20 \times 50) + (30 \times 0)} \\ &= \frac{1000 - 0}{1000 + 0} \\ &= 1 \end{aligned}$$

This index of unity — a result already anticipated — is the obvious measure of the *complete* association between maleness and delinquency, since all delinquents are male, and all females are necessarily non-delinquent.

Sensitivity of Q. Any index must be able to discriminate to some extent between slight variations in the data. It will later become apparent that each of the various measures of association has its own idiosyncra-

Table 9.2.6 *Observed and Chance Frequencies*

	Observed			*Expected*		
	B	G		B	G	
Del.	15	5	20	10	10	20
Non-del.	35	45	80	40	40	80
	50	50	100	50	50	100
	$Q = .59$			$Q = 0$		

the corresponding deficiency. There is therefore a positive affinity between boys and delinquency. However, since the excess of 15 over 10 is fairly moderate, the index reveals only a moderate association: $Q = .59$. But it is the amount of this excess that constitutes, by definition, the degree of association. This difference between the observed and expected frequencies is conventionally symbolized as $(O - E)$ — an expression that has achieved an almost epigrammatic currency in statistical methods of investigation, which the student will learn to appreciate.

Signs. Although mathematically the formula for Q will yield a sign, it may not seem very meaningful when applied to attributes such as sex, religion, or race, which do not constitute a hierarchy. In the foregoing illustration, therefore, one may say that maleness is positively associated with delinquency, but it cannot be verbalized as "the higher the sex, the higher the delinquency rate." Nevertheless, we may be guided by the algebraic sign which always refers to the association indicated in cells a and d. When the sign is positive, the joint frequencies of the attributes intersecting in cells a and d exceed chance (and are therefore linked together); when the sign is negative, the frequencies in cells a and d are less than would be expected by chance, and therefore "repel" each other. By corollary, cells b and c are read in the opposite manner. The assumption is, of course, that the alphabetical designations of cells are made in the conventional order.

Type of Data to Which Q Is Applicable. While Q is singularly appropriate for genuine attributes, nevertheless this measure may at times be applied to continuous variables by prudently cutting them into dichotomies. For example, persons may be subclassified as under and over 21 years, and responses may be either positive or negative, although these latter usually shade into each other (see Table 9.2.7).

There are, however, two cautions which should always be observed

One might intuitively suppose that the observed association of $Q = .27$ may be due in part to the fact that there are twice as many boys as girls in the total sample of 150, and that by doubling the sample of girls the index might increase. Let us, therefore, double the number of girls, retaining, however, the same *rate* of delinquency among the girls. It will be observed that the value of Q remains unchanged under this inflation. Similarly, if we double the supply of delinquents from 13 to 26 by multiplying the frequencies of that row by 2, the value of Q again remains unchanged

In general, therefore, a change in the relative size of the marginals does not affect the value of Q, so long as the ratios within either columns or rows *remain undisturbed. As we shall see later, this does not hold for* ϕ, another index of association for fourfold tables. In this respect Q is more stable than is ϕ. Whether this type of insensitivity is desirable — and should therefore be considered a virtue of stability — or whether it is a defect of the formula, must be left to the good judgment of the worker in terms of his purposes. It is doubtless important to know that the number of observations in a subclass or stratum does not, of itself, influence the size of Q.

Definition of Association: Observed and Chance Frequencies. The above illustrative tables yielded, among other things, the important generalization that, when the internal cell ratios and corresponding marginal ratios coincide, and there is no association, the index must be zero. Only when *these two sets of ratios differ, and to the extent of that difference,* is there statistical association. Indeed, the formula for Q rests on this principle of the discrepancy between the observed and chance cell frequencies.

The chance frequencies are easily derivable from the marginal frequencies according to the following logic: the boys constitute 50 per cent of the children; hence, if the two sexes are equally susceptible to delinquency (or, in other words, if the sex factor has no influence on the production of delinquency — i.e., if sex and delinquency are independent), *the boys would also have 50 per cent of the delinquencies. To express* this in still another way, if delinquencies were divided impartially (by chance) among the two sexes, the boys' quota would be 50 per cent, or in terms of our example, the boys would have 10 of the 20 delinquencies. Arithmetically put:

$$\frac{50}{100} \times 20 = 10$$

In essence, we simply adjust the cell ratios to be identical with the corresponding marginal ratios (Table 9.2.6). The boys actually show 5 more delinquencies than would be expected by chance, while the girls show

Table 9.2.8b *Unlike Marginal Sets*

50	0	50	5	40	45
25	25	50	15	40	55
75	25	100	20	80	100

competitive with other measures of association, in those situations where the cases fall predominantly in three of the four cells. Such a distribution, which is evidence of a one-way association, is possible only when marginal sets are unlike. Hence, Q is a likely choice whenever marginal sets are dissimilar. The reason for this will become clearer after ϕ and other indexes of association have been studied.

Interpretation of Yule's Q. For all its apparent simplicity and precision, Q possesses no simple quantitative meaning, and is not convertible into a specific prediction.* For example, if the association between sex and delinquency is .5, one cannot say that 50 per cent of a given sex is delinquent, or that one sex is twice as delinquent as another, or that sex accounts for 50 per cent of all delinquency. A Q-value of .8 is not twice as strong as .4, although it is stronger than all values less than .8.

Not even is a coefficient of 1.00 wholly unambiguous. In the previous example, the perfect Q did not depend on the tendency of boys to become delinquent, but rather on the relative tendency that a delinquency be committed by a boy. Actually, in absolute terms, the tendency of a boy to be delinquent could be very weak, and Q still be 1.00, provided that no girls are delinquent. A coefficient of 1.00 therefore registers the presence of one-way association, but does not reflect the degree of prediction in the other direction, although that too may be complete. We may say, therefore, that Q is primarily a measure of one-way association.

In general, it would seem to be sound practice to quote the original 2×2 tabulation along with the calculated value of Q. Not only is the table necessary in order to identify linkages between specific attributes; it also permits the reader to examine for himself the detailed structure of the relationship. But if it is a single index that is required, then one must accept the limitations of such a condensation, just as the mean was accepted as a limited representative of the full array.

* A precise probabilistic interpretation of Q has recently been formulated. See L. A. Goodman and W. H. Kruskal, "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, XLIX, 1954, p. 750.

Table 9.2.7

Opinion by Age

		Opinion Yes %	 No %
Age	−21	60	40
	21+	40	60
		100	100

in such dichotomization. (1) Such compression of a mass of continuous, detailed observations into two broad dichotomies may be an expensive waste of perfectly good data, and naturally reduces the potential precision of the resulting measure. Instead of cavalierly discarding costly accuracy, it might be advisable to use other available measures which can take such precision into account. (2) A less obvious trap for the unwary operator is the more or less arbitrary location of the cutting point, which introduces an unpredictable effect upon the ultimate index. If, in the above table, the cutting point had been set at 30 years instead of 21, the cell frequencies might have been radically changed, and the Q-measure substantially altered. Such arbitrary decisions gravely reduce *the reliability and comparability of any measure, including Q.* The student will remember the general principle that all grouping has its hazards; but when everything is staked on one cutting point, the precariousness of the undertaking is increased.

The usefulness of Q also varies with the general pattern of distribution of the data. These patterns may be particularized as follows: (1) the division of frequencies within each marginal set; (2) the similarity (dissimilarity) between the two marginal sets; and (3) the internal distribution of cell frequencies. Thus, an individual marginal set may be symmetrical (equal division) or skewed; the two marginal sets may be more or less identical in degree of symmetry; and the cell frequencies may *be approximately evenly dispersed in all four cells, concentrated in three* cells, or in two diagonal cells (see Tables 9.2.8a and 9.2.8b).

It will become increasingly evident that Q is most useful, and least

Table 9.2.8a *Like Marginal Sets*

30	25	55
25	25	50
55	50	105

5	20	25
70	5	75
75	25	100

0	50	50
50	0	50
50	50	100

(b) Convert cell frequencies to percentages of total (949), and compute *Q*.

(c) Compare answers. What principle in respect to *Q* is suggested?

Table 9.2.9 *Soldiers' Answers to Question: "Is the Army giving you a good chance to show what you can do?"*

SOLDIER CLASS	RESPONSE		TOTAL	*N*
	Yes	*No*		
Regulars	52%	48%	100%	300
Selectees	30%	70%	100%	649

Source Samuel A Stouffer *et al.*, *The American Soldier: Adjustment During Army Life*, Vol I of *Studies in Social Psychology in World War II*, Princeton University Press, Princeton, N.J., 1949, p. 73.

9. In comparing primitive tribes, a simple statistical technique is: (1) to tabulate the number of traits that two tribes have in common; (2) the number present in one but not in the other; and (3) the number of traits absent in both. Below are three such tabulations (Table 9.2.10) based on selected Indian tribes in California.
   (a) Compute *Q* for each table.
   (b) According to these results, which two tribes are culturally most similar?

Table 9.2.10 *Common Culture Traits, Selected Indian Tribes, California*

		Karok					*Karok*					*Hupa*		
		P	A				P	A				P	A	
Hupa	P	709	142	851	*Chilula*	P	527	345	872	*Chilula*	P	650	286	936
	A	238	799	1,037		A	236	848	1,084		A	82	861	943
		947	941	1,888			763	1,193	1,956			732	1,147	1,879

Source: Harold Driver, *Cultural Element Distributions: X; Northwest California*, Vol. I, No. 6, University of California Press, Berkeley, 1939, pp 309–369 (adapted).

10. Study Table 9.2.11 and answer the following questions:
    (a) If you are a smoker, what is the probability of developing lung cancer?
    (b) If you have lung cancer, what is the probability of being a smoker?
    (c) If you do not have lung cancer, what is the probability of being a smoker?
    (d) What is the probability of anyone having lung cancer?
    (e) Does the *Q*-value indicate that lung cancer is associated with smoking, or that smoking leads to lung cancer?

Questions and Problems

1. Define the following concepts:
   2×1 Table
   2×2 Table
   Statistical Association
   Statistical Independence
   One-Way Association
   Two-Way Association
   Expected Frequency
   Observed Frequency
   Marginal Distribution
   Like Marginal Sets
   Unlike Marginal Sets
   Joint Frequency Distribution

2. In a study of mental patients, "94 per cent showed evidence of status conflict before onset of mental illness." (Source: Kingsley Davis, "Mental Hygiene and Class Structure," *Psychiatry*, I, 1938, pp. 55–65.)
   (a) Does this observation demonstrate an association between status conflict and mental illness? Explain your answer.
   (b) What would be your conclusion if 94 per cent of the normal population also had status conflict?
   (c) If 50 per cent of the normal population had status conflict, what would you conclude?
   (d) Is a 2×1 table sufficient to prove a relationship?

3. Of the criminals in a given population, 80 per cent completed 8 years of schooling or less. Of the non-criminals, 20 per cent had completed more than 8 years of schooling. Prepare a 2×2 table and explain why $Q = 0$.

4. How can one know by mere inspection of a 2×2 table that Q is unity?

5. How is the sign of Q to be interpreted? Discuss with reference both to attributes and to continuous data.

6. (a) Form a complete 2×2 percentage table based on the following data: of a total of 650 young people, 80 per cent are boys, and 10 per cent are delinquents; 10 per cent of the boys are delinquent.
   (b) Convert the percentages to absolute frequencies.

7. (a) Construct a 2×2 table for the following data: 960 of 1,500 community leaders replied "Yes" to the question "If a person wanted to make a speech in your community against churches and religions, should he be allowed to speak?" Of 897 non-leaders from the same communities, 350 answered "Yes." (Source: S. A. Stouffer, *Communism, Conformity, and Civil Liberties*, Doubleday & Company, Inc., Garden City, N.Y., 1955, p. 33.)
   (b) What conclusion on the relation between leadership and tolerance is suggested by this table?

8. For Table 9.2.9:
   (a) Convert percentages to frequencies, and compute Q.

(b) Convert cell frequencies to percentages of total (949), and compute Q.
(c) Compare answers. What principle in respect to Q is suggested?

Table 9.2.9 *Soldiers' Answers to Question: "Is the Army giving you a good chance to show what you can do?"*

SOLDIER CLASS	RESPONSE		TOTAL	N
	Yes	*No*		
Regulars	52%	48%	100%	300
Selectees	30%	70%	100%	649

Source. Samuel A. Stouffer *et al., The American Soldier: Adjustment During Army Life*, Vol. I of *Studies in Social Psychology in World War II*, Princeton University Press, Princeton, N.J., 1949, p. 73.

9. In comparing primitive tribes, a simple statistical technique is: (1) to tabulate the number of traits that two tribes have in common; (2) the number present in one but not in the other; and (3) the number of traits absent in both. Below are three such tabulations (Table 9.2.10) based on selected Indian tribes in California.
(a) Compute Q for each table.
(b) According to these results, which two tribes are culturally most similar?

Table 9.2.10 *Common Culture Traits, Selected Indian Tribes, California*

		Karok P	*Karok* A	
Hupa	P	709	142	851
	A	238	799	1,037
		947	941	1,888

		Karok P	*Karok* A	
Chilula	P	527	345	872
	A	236	848	1,084
		763	1,193	1,956

		Hupa P	*Hupa* A	
Chilula	P	650	286	936
	A	82	861	943
		732	1,147	1,879

Source: Harold Driver, *Cultural Element Distributions: X; Northwest California*, Vol. I, No. 6, University of California Press, Berkeley, 1939, pp. 309–369 (adapted).

10. Study Table 9.2.11 and answer the following questions:
(a) If you are a smoker, what is the probability of developing lung cancer?
(b) If you have lung cancer, what is the probability of being a smoker?
(c) If you do not have lung cancer, what is the probability of being a smoker?
(d) What is the probability of anyone having lung cancer?
(e) Does the Q-value indicate that lung cancer is associated with smoking, or that smoking leads to lung cancer?

Table 9.2.11

Smoking and Lung Cancer

	Smokers	Non-Smokers	
Lung Cancer	9	1	10
No Lung Cancer	91	149	240
	100	150	250

Source: Hypothetical

11. In Sun County, the voters and non-voters were tabulated by sex. The association between male and voting as measured by Q was .32. In Rain County, a similar tabulation led to a Q of .64. Given these Q-values, which of the following inferences are valid? (Suggestion: prepare dummy table.)
    (a) A male in Rain County is twice as likely to vote as a male in Sun County.
    (b) The percentage of women who voted in Sun County is twice as great as the percentage of women who voted in Rain County.
    (c) In Rain County, a male is more likely to vote than a female.
    (d) In Sun County, a female is more likely to vote than a male.
    (e) What value would Q take if males and females were equally likely to vote?
    (f) A male is more likely to vote than a female in both counties.

Section Three

The Phi Coefficient (ϕ)

Comparison of ϕ and Q. The statistic ϕ (the Greek lower-case letter *phi*) has certain resemblances to Q, but in other respects it is different. Like Q, it is applicable only to 2×2 tables of true dichotomies that have no gradation in value, or to continuous variables that can be justifiably dichotomized. Similarly, ϕ measures the intensity of association which the distribution of the joint frequencies reflects. As for the formulas of ϕ and Q, the numerators are identical, but the denominators are differently constructed.

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \qquad (9.3.1)$$

where a, b, c, and d are the familiar cell frequencies, and the sums of the individual cell frequencies are the marginal subtotals. This coefficient takes the sign of the numerator, which carries the same interpretation as the sign of Q.

In order to demonstrate differences in the performance of Q and ϕ, we avail ourselves again of the cross-tabulation of delinquents by sex in which Q was unity (see **Table 9.2.3**). Applying the ϕ formula to the same data, however, we obtain a measure only half as large (Table 9.3.1).

Table 9.3.1

Delinquents by Sex

	B	G	
Del.	20	0	20
Non-del.	30	50	80
	50	50	100

$$\begin{aligned} \phi &= \frac{1000 - 0}{\sqrt{20 \cdot 80 \cdot 50 \cdot 50}} \\ &= \frac{1000}{2000} \\ &= .50 \end{aligned}$$

Q is necessarily unity, since all delinquents are boys. However, the converse, that all boys are delinquents, obviously does not hold. Hence, ϕ, whose formula is so contrived as to reflect this mutuality of relationship, is only .50. Association is *complete* but not *perfect*, or *absolute*, terms used by Yule and Kendall. Clearly, ϕ and Q do not measure the same aspects of association in the fourfold table. Essentially, the difference between the two indexes lies in the fact that the formula for ϕ is so designed that the degree of bilateral association is reflected in a single index. The ϕ coefficient "picks up" whatever two-way association there is between the two sets of attributes. Since it measures only the reciprocal relation between X and Y, it necessarily ascribes equal influence to both variables; it may therefore be said to be *reversible*.

This may be very simply illustrated (Table 9.3.2) in the case of perfect two-way association, a situation in which there is no exception in either direction: all boys are delinquent, and all delinquents are boys. As a consequence of this perfect two-way association, the joint occur-

Table 9.3.2

Perfect Two-Way Association

	B	G	
Del.	20	0	20
Non-del.	0	80	80
	20	80	100

$$\begin{aligned} \phi &= \frac{1600 - 0}{\sqrt{20 \cdot 80 \cdot 20 \cdot 80}} \\ &= \frac{1600}{1600} \\ &= 1 \end{aligned}$$

rences are restricted to one diagonal of the 2 × 2 table, and the other diagonally located cells are vacant. *Since each category completely "explains" the other,* ϕ must equal unity. To put it arithmetically, two zeros in one diagonal necessarily produce $\phi = 1$. This principle of reversibility holds good also for any intermediate value of ϕ between zero and unity when the paired *attributes only partially explain each other.* As this degree of mutuality diminishes, ϕ likewise is reduced.

Q and ϕ of course coincide in the presence of perfect two-way association for the reason that complete one-way association ($Q = 1$) is a necessary element in perfect two-way association ($\phi = 1$).

Problem of Marginal Frequencies. Like Q, ϕ of course reflects gradations in the intensity of association. Unlike Q, however, ϕ is responsive to changes in marginal ratios, *since these affect the possible degree of two-way association.* In Table 9.3.3, the sample of delinquents is only one-ninth as large as that of the non-delinquents. But when the samples of delinquents and non-delinquents are equalized, ϕ is raised from .20 to .35, whereas Q would have remained unchanged. *This sensitivity of* ϕ stems from the fact that any uniform inflation of frequencies in one row alters the ratios in column frequencies, both marginal and internal, and thereby alters (in this instance, improves) the potentiality of a mutual *relationship.* Q *and* ϕ *are left unaltered if the whole table is multiplied* through by a constant (Table 9.3.3); in consequence, both ϕ and Q can be calculated from percentages to the base N.

Table 9.3.3 *Effect of Marginal Distributions*

	B	G		B	G		B	G	
Del.	8	2	10	72	18	90	16	4	20
Non-del.	42	48	90	42	48	90	84	96	180
	50	50	100	114	66	180	100	100	200
	$\phi = .20$			$\phi = .35$			$\phi = .20$		

This sensitivity of the index to shifts in marginal ratios may appear to be a defect of the formula. If a pair of subtotals exerts such an influence on the value of ϕ, a statistical worker could rig his coefficient up or down by tampering with subtotals. Moreover, two or more otherwise reliable studies may lack comparability merely because of the uncontrolled factor of subsample size.

Since marginals are occasionally subject to quite arbitrary choice, and sometimes are necessarily quite unbalanced, one may inquire whether there is a "natural" or "ideal" marginal ratio. Is perhaps equality in size of paired samples to be considered the most valid?

This may at first glance appear to be an inviting prospect, since the resulting 50–50 division of the marginal subtotals would theoretically allow the cell frequencies to vary maximally, and thereby yield every possible gradation of ϕ. But since many sociologically interesting dichotomies (e.g., delinquent and non-delinquent, married and divorced) exist in nature only in disproportionate supply, it would be misleading to set up equal subtotals. And even if an equal division in attributes were arbitrarily set up for the independent variable, the dependent variable would still have to be allowed to vary as it may. Thus, in much social inquiry, there is no escape from unequal, and even disproportionate, subsamples. In those instances, if the worker has any misgivings about the validity of his obtained ϕ-coefficient, he can always quote his entire tiny table for the information of his reader.

ϕ of Unity Possible Only with Identical Marginal Sets. It has already been stated that all correlation formulas are ideally devised so as to permit the indexes to vary between zero and unity. Any circumstance which limits the range within which the coefficient can vary blemishes the norming function of the coefficient, thereby reducing the usefulness of the index. It is as though percentage distributions were not permitted to vary between 0 and 100. The ϕ-coefficient is in exactly that dilemma when marginal sets are not identical. Under these conditions, no amount of juggling cell frequencies (Table 9.3.4) will make it possible to enter zeros in two diagonal cells while maintaining intact the given observed marginals.

Table 9.3.4 *Unlike Marginal Sets and ϕ_{max}*

<table>
<tr><td>8</td><td>2</td><td>10</td><td>7</td><td>3</td><td>10</td><td>10</td><td>0</td><td>10</td><td>5</td><td>5</td><td>10</td></tr>
<tr><td>12</td><td>3</td><td>15</td><td>13</td><td>2</td><td>15</td><td>10</td><td>5</td><td>15</td><td>15</td><td>0</td><td>15</td></tr>
<tr><td>20</td><td>5</td><td>25</td><td>20</td><td>5</td><td>25</td><td>20</td><td>5</td><td>25</td><td>20</td><td>5</td><td>25</td></tr>
<tr><td colspan="3">$\phi = 0$</td><td colspan="3">$\phi = -.20$</td><td colspan="3">$\phi = .41$</td><td colspan="3">$\phi_{max} = -.61$</td></tr>
</table>

Although identity of marginal sets is necessary for ϕ to be unity, it is not sufficient of itself to guarantee perfect correlation. Identical marginals merely *permit* the value of ϕ to range within the ideal limits. It is always the *internal* ratios which determine the correlation index, as is apparent in Table 9.3.5.

rences are restricted to one diagonal of the 2×2 table, and the other diagonally located cells are vacant. Since each category completely "explains" the other, ϕ must equal unity. To put it arithmetically, two zeros in one diagonal necessarily produce $\phi = 1$. This principle of reversibility holds good also for any intermediate value of ϕ between zero and unity when the paired attributes only partially explain each other. As this degree of mutuality diminishes, ϕ likewise is reduced.

Q and ϕ of course coincide in the presence of perfect two-way association for the reason that complete one-way association $(Q = 1)$ is a necessary element in perfect two-way association $(\phi = 1)$.

Problem of Marginal Frequencies. Like Q, ϕ of course reflects gradations in the intensity of association. Unlike Q, however, ϕ is responsive to changes in marginal ratios, since these affect the possible degree of two-way association. In Table 9.3.3, the sample of delinquents is only one-ninth as large as that of the non-delinquents. But when the samples of delinquents and non-delinquents are equalized, ϕ is raised from .20 to .35, whereas Q would have remained unchanged. This sensitivity of ϕ stems from the fact that any uniform inflation of frequencies in one row alters the ratios in column frequencies, both marginal and internal, and thereby alters (in this instance, improves) the potentiality of a mutual relationship. Q and ϕ are left unaltered if the whole table is multiplied through by a constant (Table 9.3.3); in consequence, both ϕ and Q can be calculated from percentages to the base N.

Table 9.3.3 *Effect of Marginal Distributions*

	B	G		B	G		B	G	
Del.	8	2	10	72	18	90	16	4	20
Non-del.	42	48	90	42	48	90	84	96	180
	50	50	100	114	66	180	100	100	200
$\phi = .20$				$\phi = .35$			$\phi = .20$		

This sensitivity of the index to shifts in marginal ratios may appear to be a defect of the formula. If a pair of subtotals exerts such an influence on the value of ϕ, a statistical worker could rig his coefficient up or down by tampering with subtotals. Moreover, two or more otherwise reliable studies may lack comparability merely because of the uncontrolled factor of subsample size.

Since marginals are occasionally subject to quite arbitrary choice, and sometimes are necessarily quite unbalanced, one may inquire whether there is a "natural" or "ideal" marginal ratio. Is perhaps equality in size of paired samples to be considered the most valid?

This may at first glance appear to be an inviting prospect, since the resulting 50–50 division of the marginal subtotals would theoretically allow the cell frequencies to vary maximally, and thereby yield every possible gradation of ϕ. But since many sociologically interesting dichotomies (e.g., delinquent and non-delinquent, married and divorced) exist in nature only in disproportionate supply, it would be misleading to set up equal subtotals. And even if an eqnal division in attributes were arbitrarily set up for the independent variable, the dependent variable would still have to be allowed to vary as it may. Thus, in much social inquiry, there is no escspe from unequal, and even disproportionate, subsamples. In those instances, if tbe worker has any misgivings about the validity of his obtained ϕ-coefficient, he can always quote his entire tiny table for the information of his reader.

ϕ of Unity Possible Only with Identical Marginal Sets. It has already been stated that all correlation formulas are ideally devised so as to permit the indexes to vary between zero and unity. Any circumstance which limits the range within whicb the coefficient can vary blemishes the norming function of the coefficient, thereby reducing the usefulness of the index. It is as though percentage distributions were not permitted to vary between 0 and 100. The ϕ-coefficient is in exactly that dilemma wben marginal sets are not identical. Under these conditions, no amount of juggling cell frequencies (Table 9.3.4) will make it possible to enter zeros in two diagonal cells while maintaining intact the given observed marginals.

Table 9.3.4 *Unlike Marginal Sets and ϕ_{max}*

8	2	10	7	3	10	10	0	10	5	5	10
12	3	15	13	2	15	10	5	15	15	0	15
20	5	25	20	5	25	20	5	25	20	5	25
$\phi = 0$			$\phi = -.20$			$\phi = .41$			$\phi_{max} = -.61$		

Although identity of marginal sets is necessary for ϕ to be unity, it is not sufficient of itself to guarantee perfect correlation. Identical marginals merely *permit* the value of ϕ to range within the ideal limits. It is always the *internal* ratios which determine the correlation index, as is apparent in Table 9.3.5.

Table 9.3.5 *Identical Marginal Sets, Varied Internal Ratios*

113	37	150	100	50	150	140	10	150	150	0	150
37	13	50	50	0	50	10	40	50	0	50	50
150	50	200	150	50	200	150	50	200	150	50	200
$\phi = .01$			$\phi = -.33$			$\phi = .73$			$\phi = 1$		

Calculation of Maximum ϕ. Where a ϕ of unity is impossible to attain, it may nevertheless seem desirable to determine the maximum possible ϕ for the given unlike marginal sets, and this maximum ϕ may then be set up as a standard against which to assess the obtained value. The maximum value of ϕ may be obtained by placing a zero in one cell, which guarantees complete one-way association, and then arranging the other frequencies so as to maximize the divergence between *ad* and *bc*. Table 9.3.4 exemplifies this exploratory procedure.

However, instead of resorting to such experimental shuffling, ϕ_{max} can be readily computed for any given marginals by means of the following formula:

$$\phi_{max} = \sqrt{\frac{s_i}{l_i} \times \frac{l_j}{s_j}} \tag{9.3.2}$$

in which l and s = the larger and smaller individual subtotals of the respective marginal sets

l_i = the largest subtotal in the table

l_j = the larger subtotal in the other set

Applying this formula to the original distribution of Table 9.3.4, we find that:

$$\phi_{max} = \sqrt{\frac{5}{20} \times \frac{15}{10}} = \pm .61$$

This numerical result corresponds to the ϕ-value of $-.61$ obtained by trial and error.

The ϕ formula does not pose the same problem on the zero end of the coefficient range. Although ϕ may show a variable maximum, the possible minimum is always zero. Any set of marginal ratios, irrespective of how unbalanced or unlike, is compatible with a zero correlation. The reason is that any marginal ratio can be theoretically matched within columns and rows (Table 9.3.6) so as to yield $\phi = 0$. Hence, we have no occasion to speak of ϕ_{min}, but only of ϕ_{max}.

Table 9.3.6 *Phi*$_{min}$ = *Zero*

25	25	50	50	50	100	10	25	35
75	75	150	50	50	100	20	50	70
100	100	200	100	100	200	30	75	105
$\phi = 0$			$\phi = 0$			$\phi = 0$		

Function of ϕ_{max}. The fact that ϕ can attain a maximum of unity only when marginal sets are identical strips it of a portion of its norming capacity, for two or more coefficients are not comparable if they happen to be based on diverse marginals. It has been suggested that an escape from this dilemma is to express the observed ϕ as a proportion of the maximum possible ϕ. But such a figure still does not convey any conception of the degree of association. By that technique, an observed ϕ of .2 and a ϕ_{max} of .2 would yield a "corrected" ϕ of unity! This would certainly be an overstatement of the existing association.

In any event, it is never clear whether a low value of ϕ is due to the inhibiting force of the marginal frequencies or to a weak intrinsic relation between the variables as evidenced by cell frequencies.

Type of Data to Which ϕ *Is Applicable.* In the first place, we assume that the categories are true dichotomies, such as race, religion, employment status, or other categories which show no gradation in value. However, continuous data may be compressed into dichotomies and treated like a pair of attributes, as has been previously suggested. There is nothing in the mathematics of the ϕ formula which forbids such utilization of data, for abstract mathematics is quite oblivious to the meaning of the substantive empirical data to which the measures refer. The real problem lies in the interpretation of the data.

Secondly, the distribution of marginal frequencies must be taken into account in judging the appropriateness of ϕ. Since there is a question of the fundamental usefulness of ϕ when ϕ_{max} is appreciably lower than unity, ϕ should be very cautiously applied under these circumstances; and this for two reasons: (1) when ϕ is applied to unbalanced sets of marginals, in which two-way association is negligible, it will necessarily understate the one-way relationship; and (2) a series of ϕ-values with varied maxima cease to be comparable. This predicament is serious, since comparability is one of the basic objectives of norming in the first place.

What renders the use of ϕ all the more problematical for application to social data is the fact that matched marginals occur very rarely in nature, as has already been implied. We would therefore either have to abandon the employment of ϕ, except in such rare cases as those in which identical marginals were present, or tolerate an approximation when the discrepancy is not too great and the requirements of the problem not too stringent. Practically speaking, this is a common and justifiable solution of such a dilemma, since statistics thrives on approximations.

Uses of ϕ and Q Compared. Since a large proportion of sociological research deals with attributes, the fourfold table would seem very useful to sociologists. And in the event that it is desirable to express association in terms of either Q or ϕ, the following guiding principles may serve as rough criteria of choice between them:

(1) When the marginal sets are very unlike, it will be reasonable to use Q in order to salvage whatever one-way association there is.
(2) When marginal sets are approximately alike, use ϕ in order to test the presence of two-way association.
(3) When ϕ is low for any reason, try Q.

In spite of its limitations, the ϕ-coefficient, as distinguished from Q, possesses certain characteristics which should not be overlooked and tend to explain its current preferential use over Q. First, its derivation is statistically more rigorous than that of Q; second, it is mathematically equivalent to the Pearsonian product-moment correlation (Chapter 10, Section 3); and third, it is functionally related to chi-square (Chapter 9, Section 4). However, in spite of a certain elegance, ϕ is hampered — as is, for that matter, every other index — by certain characteristics which limit its practical utility. For all its neatness, it still cannot take over the functions which the less useful Q can perform better.

Questions and Problems

1. While in training in the United States (1943) and later in Europe (1945), the same group of 100 soldiers were asked: "In general, how do you feel most of the time, in good spirits or in low spirits?" Their replies are shown in Table 9.3.7:

Table 9.3.7

Soldier Morale, U.S. (1943) and Europe (1945), Percentage Distribution

1945	*1943*		
	Good Spirits	Low Spirits	
Good Spirits	27	17	44
Low Spirits	9	47	56
	36	64	100

Source: Samuel A. Stouffer *et al.*, *The American Soldier: Adjustment During Army Life*, Vol. I of *Studies in Social Psychology in World War II*, Princeton University Press, Princeton, N.J., 1949, p. 163.

(a) What percentage of the total group responded in the same way on both occasions?
(b) Compute ϕ.
(c) What percentage of each group changed their response?
(d) Interpret these results: compare the information which the respective answers yield.

2. The following question was put to 624 high school boys: "Which boy in the senior class seems to you the most poised in social situations?" Their votes, classified by religious background of choosers and chosen, are shown in Table 9.3.8a.

Table 9.3.8a

Sociometric Choices by Religious Background

	Chosen		
Chooser	Jews	Non-Jews	
Jews	239	44	283
Non-Jews	77	264	341
	316	308	624

Source: Jackson Toby, "Universalistic and Particularistic Factors in Role Assignment," *American Sociological Review*, XVIII, 1953, p. 134.

(a) Compute ϕ and ϕ_{max}.
(b) What does Table 9.3.8a suggest about the effect of religious ethnocentrism?
(c) What proportion of the respective choosers chose their own social group?
(d) What proportion of the total chose their own group?

3. The votes for the ten most popular boys (chosen most frequently) were classified by religion of choosers and chosen, as shown in Table 9.3.8b.

Table 9.3.8b

Sociometric Choices by Religious Background, Ten Most Popular Boys

	Chosen		
Chooser	Jews	Non-Jews	
Jews	124	25	149
Non-Jews	57	55	112
	181	80	261

(a) What proportion of the total (261) chose their own group?
(b) What proportion of each group chose their own group?

4. All other boys were classified similarly (Table 9.3.8c).

Table 9.3.8c

Sociometric Choices by Religious Background, Boys Not Among First Ten

Chooser	*Chosen* Jews	Non-Jews	
Jews	115	19	134
Non-Jews	20	209	229
	135	228	363

(a) Compute ϕ for Tables 9.3.8b and 9.3.8c.

(b) Would these results modify the conclusion based on Table 9.3.8a?

5. In an investigation of nonliterate cultures it is found that 19 out of 34 agricultural tribes, and 4 out of 24 non-agricultural tribes practice slavery.

Table 9.3.9 *Slavery by Type of Economy, 58 Primitive Cultures*

	Agricultural	*Non-agricultural*	*Total*
Slavery	19	4	23
No slavery	15	20	35
Total	34	24	58

Source: Leo W. Simmons, "Statistical Correlations in the Science of Society," in G. P. Murdock, *Studies in the Science of Society, Presented to Albert G. Keller*, Yale University Press, New Haven, Conn., 1937, Table 2.

(a) Which index, Q or ϕ, seems more appropriate to this tabulation?

(b) Which variable would you consider dependent?

(c) Which attributes afford the best prediction of the others?

(d) Prepare percentage distributions in each direction and compare.

6. (a) Compute Q and ϕ for Table 9.3.10.

(b) Why do they differ so greatly? Is Q preferable?

(c) Compute the maximum ϕ and compare with observed ϕ.

(d) Discuss the adequacy of the following summaries of the tabulation:

Color and occupation are correlated.

Color and occupation are positively correlated.

The correlation between color and occupation is +.29.

The correlation between color and occupation, as measured by Q, is .8.

More non-white women are employed in domestic service than white women.

Being non-white is associated with being employed as a domestic.

A larger proportion of non-white women are in domestic service than white women.

In 409,000 employed women in Chicago, 1940, being non-white is associated with being employed in domestic service and the degree of that association, as measured by 'Q, is .8.

Table 9.3.10 *Type of Occupation by Race, Employed Females (in '000), Chicago, 1940*

	Occupation		
Color	Other than Domestic Service	Domestic Service	*Total*
White	359	23	382
Non-white	17	10	27
Total	376	33	409

Source: O. D. Duncan and Beverly Davis, "An Alternative to Ecological Correlation," *American Sociological Review*, XVIII, 1953, p. 665.

7. Investigators interested in political attitudes asked a group of people the following question: "If you had to choose for president between a man who has had experience in government and a man who has had experience in business, which would you choose?" Table 9.3.11 presents a summary of the replies classified by major political parties.

Table 9.3.11 *Preferred Occupational Background of President by Major Political Parties, 1948*

	Party of Respondent		
Response	Democratic	Republican	*Total*
Business	34	161	195
Government	115	47	162
Total	149	208	357

(c) Discuss in your own words the association between party affiliation and type of person preferred for president.

8. In general, which seems to be more informative, the analysis of rows and columns, or the comprehensive index?

SECTION FOUR

Coefficient of Contingency

Yule's Q and ϕ apply only to dichotomous classifications, hut it is often not only undesirable, but even impossible, to reduce sociological data to a dichotomy. Although male–female, insane–sane, yes–no data may logically fall into a 2×2 table, the equally prevalent socio-economic classes, religious denominations, types of crime, gradations of attitude, and many other social groupings require more than two categories. To cross-tabulate such *polytomous* classifications, a 2×2 contingency table is insufficient; a *manifold table* is required.

We shall consider a formula which answers the requirement of such manifold distributions, namely, Karl Pearson's *coefficient of contingency*, usually abbreviated to C. It is based on the deviations of the observed cell frequencies from those frequencies expected on the assumption of chance, as measured by χ^2 (read "chi-square").

Chi-Square. The statistic χ^2 has a rather elaborate mathematical grounding, and yet it is a very conventional quantity, useful in those statistical situations where it is necessary to measure the discrepancy between the observed and expected frequencies. Here, however, it will be compactly set forth for practical purposes without fundamental explanation, since it involves principles which may not he readily comprehensible at this stage of the student's development. The steps in the computation of χ^2 are as follows:

(1) Compute the expected frequencies (E).
(2) Subtract the expected from the observed frequencies ($O - E$).
(3) Square each difference $(O - E)^2$.
(4) Divide each squared discrepancy by its expected frequency, $\frac{(O-E)^2}{E}$, thereby norming each absolute discrepancy on its own base.
(5) Sum the resulting ratios, $\sum \frac{(O-E)^2}{E}$; this sum is, by definition, χ^2.

The formula will therefore read:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad (9.4.1)$$

Given a joint frequency table such as the following, the calculation of χ^2 would proceed as shown in Table 9.4.1.

18	2	20
6	34	40
24	36	60

Computation of Expected Frequencies. The expected frequencies are computed in exactly the same manner as has already been explained in Section 2 of this chapter. As was then defined, chance frequencies in rows and columns are proportional to corresponding marginal totals. In the above example, the proportion of all items in the first row is ⅓; hence, the chance frequencies in the first row would be one-third of the respective column totals. Thus, the expected frequency in Row 1, Column 1, is (⅓) × (24) = 8. One could compute the other expected frequencies by subtraction from the marginal subtotals. A prudent check on accuracy, however, can be secured by independently computing all cell frequencies in the aforesaid manner, and totaling them to determine whether they correspond to the marginal totals. The expected frequency of any cell may be routinely calculated by multiplying its marginal frequencies together and dividing by N. In symbols:

Table 9.4.1 *Computation of χ^2, 2 × 2 Contingency Table, Worksheet*

Intersection of: Row	and Col.	O	E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
1	1	18	8	10	100	12.50
1	2	2	12	−10	100	8.33
2	1	6	16	−10	100	6.25
2	2	34	24	10	100	4.17
Total		60	60	0		$\chi^2 = 31.25$

$$E_{ij} = \frac{r_i c_j}{N}$$

where r_i = marginal frequency of ith row
c_j = marginal frequency of jth column

Coefficient of Contingency. The coefficient of contingency is designed to measure the degree of contingency, or dependence, between two variables or sets of attributes. Now, since pure chance distribution indicates no association at all, the more nearly this pure chance distribution approximates the observed distribution, the weaker the affinity must necessarily be, similarly, the greater the discrepancy between the observed and chance distributions, the greater must be the association, or dependence, between the variables. Since χ^2 has been chosen as the measure of this discrepancy, the higher the χ^2-value, the greater the association; χ^2 could therefore be accepted as a rough measure of correlation.

However, it cannot qualify as a standard correlation measure, since its upper limit varies directly with the number of observations, N, so that successive instances of χ^2 will be lacking in comparability. In other words, raw χ^2-values are not normed; they do not range from zero to unity, and therefore are not suited for a measure of correlation in the accepted sense.

But the following formula, which allows for varying N's, more or less satisfies this requirement of a standard range:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{9.4 2}$$

It will be clear upon examination of the formula that if χ^2 is large in relation to N, C will approach unity, since numerator and denominator will be virtually equal; however, if χ^2 is small in relation to N, the coefficient will also be small and tend towards zero. If $\chi^2 = 0$ (that is, if there is no discrepancy between the observed data and pure chance), the coefficient will, of course, also be zero because the numerator is zero.

Demonstration of C. In order to comprehend the type of association measured by the contingency coefficient, let us set up a hypothetical case of obviously perfect correlation between class membership and political opinion which is not disturbed by exceptions, and a similar tabulation which is not perfect. In Table 9.4.2a, all members of the highest class are conservative, and all conservatives are in the highest class without exception in either direction — and so on through the remaining categories. The absence of exceptions to this generalization is indicated by the six zero entries. No greater perfection of association can be conceived, no matter what the marginal distributions, so long as they are identical. As members of the social classes move down the scale, they also move uniformly across to the radical end of the opinion scale.

However, in Table 9.4.2b, not all of the frequencies fall on the diagonal; but the dominant trend along the diagonal may still be discerned, since the larger frequencies are concentrated there. Nevertheless, a considerable number of persons diverge from the observed trend. In Table 9.4.2a the prediction that a person of Class I would be conservative

Table 9.4.2a *Computation of Contingency Coefficient, Perfect Association, Political Opinion by Social Class*

Social Class	Observed				Expected			
	Political Attitude:			Total	*Political Attitude:*			Total
	Cons.	Neut.	Rad.		Cons.	Neut.	Rad.	
High	13	0	0	13	3.8	4.3	4.9	13
Middle	0	15	0	15	4.3	5.0	5.7	15
Low	0	0	17	17	4.9	5.7	6.4	17
Total	13	15	17	45	13	15	17	45

$$C = \sqrt{\frac{90}{90 + 45}} = .816$$

Source: Hypothetical

would be correct 100 per cent of the time. On the other hand, in Table 9.4.2b, such a prediction would be much less certain, since it is correct only 10 out of 15 times. The coefficient of contingency would accordingly be lower, and is found to be .57.

Formula Understates Degree of Association. But C does not function perfectly according to convention. The formula is not so constructed that it can successfully yield unity, even though the intrinsic relation between the attributes be perfect. Thus, in spite of the evident perfect correlation in Table 9.4.2a, the obtained coefficient was only .816 instead of unity. To increase the complications, this maximum obtainable index (C_{max}), the equivalent of perfect correlation, varies with the number of cells in the table.

It may be shown that we can calculate the theoretical C_{max} on the sole basis of the number of rows (or columns), provided the table is square and marginal sets are identical. For square tables, we find that:

$$C_{max} = \sqrt{\frac{t-1}{t}} \qquad (9.4.3)$$

where t = the number of rows

Table 9.4.2b *Computation of Contingency Coefficient*

SOCIAL CLASS	OBSERVED				EXPECTED			
	Political Attitude:			TOTAL	*Political Attitude:*			TOTAL
	Cons.	Neut.	Rad.		Cons.	Neut.	Rad.	
High	10	3	2	15	5.6	4.4	5.0	15
Middle	7	8	1	16	6.0	4.7	5.3	16
Low	2	4	14	20	7.4	5.9	6.7	20
Total	19	15	17	51	19	15	17	51

Row and Col.		O	E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
1	1	10	5.6	4.4	19.36	3.46
1	2	3	4.4	−1.4	1.96	.45
1	3	2	5.0	−3.0	9.00	1.80
2	1	7	6.0	1.0	1.00	.17
2	2	8	4.7	3.3	10.89	2.32
2	3	1	5.3	−4.3	18.49	3.49
3	1	2	7.4	−5.4	29.16	3.94
3	2	4	5.9	−1.9	3.61	.61
3	3	14	6.7	7.3	53.29	7.95
		51	51.0	0.0		$\chi^2 = 24.19$

$$C = \sqrt{\frac{24.19}{24.19 + 51}} = \sqrt{.3217} = .57$$

Source: Hypothetical

perfect, provided marginal sets are identical. The ratio between the obtained and the maximum C would therefore be roughly equivalent to the conventional measure of association, ranging from 0 to 1. Applying this reasoning to Table 9.4.2a where the obtained C was .816, we would compute the adjusted C:

$$C_{adj} = \frac{C}{C_{max}}$$
$$= \frac{.816}{.816}$$
$$= 1$$

which is obviously correct.

This adjustment becomes smaller and smaller, and therefore decreasingly necessary, as the dimensions of the table increase.

The Sign of C. The *C*-coefficient normally carries no sign, and this for two simple reasons. (1) Since *C* is the square root of a number, it may be either plus or minus; hence, no definite sign is imposed on the index on strictly mathematical grounds. (2) In polytomous tables, the joint frequencies need not be restricted to the diagonal cells, even though the correlation is perfect; hence, the sign is meaningless. However, if the categories may be *ordered* from high to low, there is no objection to establishing the sign by inspection. Thus, in the present example, we may speak of a high positive correlation between class status and degree of political conservatism, since as class standing rises, intensity of conservatism also increases.

Conditions Under Which C Is Appropriate. (1) As in the case of *Q* and ϕ, *C* is designed to measure association between qualitative variables. (2) The correlation between two series of data is most easily verbalized if the categories can be ordered (formulated in graded sequence): for example, social classes ranked from high to low, or social attitudes ranked from most favorable to least favorable. This ordering, however, is not a mathematical requirement. Since χ^2 is merely the sum of all the normed discrepancies between observed and expected frequencies, their order is technically immaterial. Furthermore such categories as race and religious denomination do not usually lend themselves to natural ordering, but they are not thereby excluded from consideration for this type of correlation. In such an instance, the index loses nothing in validity, although the interpretation is likely to be awkward and verbose. (3) Because of the fact that C_{max} approaches unity only when the number of cells is large, it is sometimes recommended that *C* be computed only for square tables of at least 5×5 dimension.

As originally formulated by Karl Pearson, the *C*-coefficient for attributes was put forward as an approximation of the product-moment correlation coefficient for continuous variables. However, the demanding assumption underlying this equivalence — normally distributed variables which assure a linear relationship — is almost prohibitive for sociological data.

The general trend is, therefore, toward a broadening of its scope of application beyond that originally conceived by its inventor. What *C* thereby loses in precision of interpretation, it gains in resourcefulness. It is commonly employed without regard to assumptions of linearity or of normality of marginal distributions in rectangular tables of continuous, discrete, or qualitative data. Its limitations are evident, however, from the fact that only square tables can yield a perfect correlation (otherwise marginals cannot be identical); and indexes drawn from unlike tables are not comparable.

Table 9.4.2b *Computation of Contingency Coefficient*

Social Class	Observed *Political Attitude:* Cons.	 Neut.	 Rad.	 Total	 *Politica'* Cons.
High	10	3	2	15	5.6
Middle	7	8	1	16	6.0
Low	2	4	14	20	7.4
Total	19	15	17	51	19

Row and Col.		*O*	*E*	*O − E*
1	1	10	5.6	4.4
1	2	3	4.4	−1.4
1	3	2	5.0	−3.0
2	1	7	6.0	1.0
2	2	8	4.7	3.3
2	3	1	5.3	−4.3
3	1	2	7.4	−5.4
3	2	4	5.9	−1.0
3	3	14	6.7	7.3
		51	51.0	0.0

$$C = \sqrt{\frac{24.19}{24.10 + 51}} .$$

Source: Hypothetical

perfect, provided marginal sets ar
tained and the maximum C wo
the conventional measure of asso
this reasoning to Table 9.4.2a w
compute the adjusted C:

$$C_{*}$$

which is obviously correct.

This adjustment becomes small
ingly necessary, as the dimensions

Table 9.4.4 *Marital Adjustment of Husbands by Degree of Attachment to Father*

Degree of Attachment	*Marital Adjustments*			
	Poor	Fair	Good	*Total*
Little or none	32 (20)	28 ( )	15 ( )	75
Moderate	41 (43)	47 ( )	69 ( )	157
A good deal	26 (35)	41 ( )	61 ( )	128
Very close	28 ( )	22 ( )	59 ( )	109
Total	127	138	204	469

Source. Ernest W. Burgess and Leonard S. Cottrell, Jr., *Predicting Success and Failure in Marriage*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1939, p. 377.

Selected References

Goodman, Leo, and William H. Kruskal, "Measures of Association for Cross Classification," *Journal of the American Statistical Association*, XLIX, 1954. Pages 732-763.

Guttman, Louis, "The Qualitative Prediction of a Qualitative Variate," in *The Prediction of Personal Adjustment*, edited by Paul Horst. Social Science Research Council, New York, 1941. Pages 258-263.

Kendall, Maurice G., *The Advanced Theory of Statistics*, 5th edition. Hafner Publishing Co., New York, 1952. Volume 1, Chapter 13.

Yule, G. Udny, and Maurice G. Kendall, *An Introduction to the Theory of Statistics*, 14th edition. Hafner Publishing Co., New York, 1950. Chapter 2.

Yule, G. Udny, *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, England, 1944. Chapters 7 and 8.

Zelditch, Morris, Jr., *A Basic Course in Sociological Statistics*. Henry Holt and Company, New York, 1959. Chapter 7.

unless the association is perfect. As in so many other situations, we must be content with approximations

Questions and Problems

1. Define the following concepts:
   Coefficient of Contingency
   Contingency Table
   Polytomous Classification
   Chi-Square
   $C_{\max}$

2. (a) Compute C for Table 9.4.3.
   (b) Why is C difficult to interpret for this table?

Table 9.4.3 *Peer Ratings by Social Class, High School Students*

Peer Rating	*Social Class* I and II	III	IV	V	*Total*
Elite	27	30	9	0	66
Good Kids	8	114	133	4	259
Grubby Gang	0	2	41	22	65
Total	35	146	183	26	390

Reprinted with permission from A. B. Hollingshead, *Elmtown's Youth*, p 222. Copyright 1949. John Wiley & Sons, Inc.

3. (a) Calculate the remaining expected frequencies in Table 9.4.4.
   (b) Compute C.
   (c) Describe verbally the relationship between variables.
   (d) Compute $C_{\max}$ by combining degrees of attachment (little and moderate) for a square table. Compare with C.

4. Compute $C_{\max}$ for 4×4 and 8×8 tables.

5. Compute Q, ϕ, and C for the following 2×2 table. Compare results and discuss the differences.

8	2	10
2	8	10
10	10	20

rank order is based on the concept of "more or less," even though the "unit of measure" is subjective, we may conceive of the collective items ranked as a quantitative, rather than qualitative variable; and when two sets of ranks vary together, we speak of *rank-order correlation.*

Rank-Difference Method of Correlation (ρ). The simplest correlation of ranks is that between only two sets of ranks. When the number of items ranked are no more than a half dozen, an observer might by inspection be able to draw a rough but acceptable conclusion on the degree of correlation between them.

Consider the two series of rankings of six pictures by two judges shown in Table 10.1.1a. Because the two rankings are duplicates, the discrepancies between the paired rankings are all zero. Furthermore, there is perfect prediction from one rank order to the other. The measure of correlation between ranks would therefore logically be unity.

Table 10.1.1a

Rank-Order Correlation: Six Items, Two Judges; Perfect Positive Correlation

Picture	Judge X	Judge Y	Difference (*D*)
A	1	1	0
B	2	2	0
C	3	3	0
D	4	4	0
E	5	5	0
F	6	6	0

On the other hand, one set of ranks could be arranged from low to high (Table 10.1.1b) instead of from high to low without affecting the degree of predictability. While the two series of ranks are now in exactly reverse order, the deviation of each rank relative to the mean rank is unchanged, and therefore mutual predictability remains unchanged. Hence, the measure of relation is still unity, but now negative.

Table 10.1.1b

Perfect Negative Correlation

Picture	*X*	*Y*	*D*	D^2
A	1	6	−5	25
B	2	5	−3	9
C	3	4	−1	1
D	4	3	1	1
E	5	2	3	9
F	6	1	5	25
Total			0	70

Covariation: Quantitative Variables 10

SECTION ONE

Correlation of Ranks: Rho

Concept of Covariation. Heretofore, correlation procedures were restricted to measuring the relationship among qualitative variables on the basis of joint frequencies. But quantitative variables, too, may be associated, and are then said to *covary:* a change in one variable is paralleled by a change in the other. Birth rates and family income, or suicide rates and proportion of Protestants, may be linked together, and the degree of association measured by an appropriate index of correlation. Of the many correlation indexes which have been devised, only three will here be elaborated, each corresponding to a given type of data and pattern of relationship: (1) paired ranks (rho), (2) linearly related series (r), and (3) curvilinearly related series (eta).

Measurement by Ranks. A very simple ordering of data is in the form of ranks (Chapter 2, Section 1). The simplicity consists in the fact that measurement may be intuitive and subjective, rather than by palpable, objective units of measure. Such rankings may not be very precise, and yet they may be very useful. In fact, many sociological concepts cannot be conveniently manipulated in any other manner. Consequently, a considerable volume of social research is founded on such a subjective base. Thus, occupations may be ranked by prestige; fellow students may be ranked in order of preference; races and nationalities may be ranked by favorable or unfavorable prejudice; pictures may be ranked by aesthetic value. Even when units of measure are available — as, for example, in the sizes of cities or the magnitudes of birth rates — ranking may be resorted to when such precision is not required. Since

A third possible relationship — between these two extremes — would be the chance relation, in which the rank of X cannot be predicted from the rank of Y (Table 10.1.1c): any X-rank is equally likely to be paired with any Y-rank. In this case, at least theoretically, the correlation would be zero.

Table 10.1.1c

Chance Relation

Picture	X	Y	D	D^2
A	1	3	−2	4
B	2	5	−3	9
C	3	1	2	4
D	4	6	−2	4
E	5	2	3	9
F	6	4	2	4
Total			0	34

A formula to measure the degree of correlation between two sets of ranks, and normed to yield the conventional range of zero to unity, was derived by Charles Spearman in 1904:

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \qquad (10.1.1)$$

where D = difference between paired ranks
N = number of items ranked

Solving for the three examples just given, we obtain the followiog coefficients:

Table 10.1.1a

$$\begin{aligned}\rho &= 1 - \frac{0}{6(36-1)}\\ &= 1\end{aligned}$$

Table 10.1.1b

$$\begin{aligned}\rho &= 1 - \frac{6(70)}{6(35)}\\ &= 1 - 2\\ &= -1\end{aligned}$$

Table 10.1.1c

$$\begin{aligned}\rho &= 1 - \frac{6(34)}{6(35)}\\ &= 1 - .97\\ &= .03\end{aligned}$$

Procedure in Case of Tied Ranks. Ranks occasionally will be tied. A puzzled judge may rate two pictures identically; quantitative scores and measures may also tie. In such cases two or more items may seem to share the same rank. Since the number of ranks and the number of items must coincide, it will simply not be possible for two items to occupy the same rank, so they must be given adjoining ranks. In such instances, both items will be assigned the arithmetic average of the two adjoining ranks. Thus, if the third and fourth items are tied, each is given the rank of 3.5. If three or more items are tied, the same rule of averaging judges ranking the same exhibit of pictures may be at wholly different locations on the complete continuum of taste preference, but still rank the art works in identical order. Judge I may regard them all as on a high level of excellence, while Judge II may see them collectively as a poor lot. The judges may agree on the order, but disagree on the qualities themselves. Nevertheless, in many actual situations, it is reasonable to suppose that similar rankings do correspond to similar preferences, owing to the fact that by and large individuals share a common culture. When, however, the two series are not on a single continuous scale, as in the case of rental and income, or when they are made up of incommensurable categories, as, for example, birth rates and income, this interpretative problem of scale location does not, of course, arise. Still and all, the worker should be certain, before employing ρ, that he is interested in correlating only rank orders rather than actual magnitudes.

A Measure of Agreement in Three or More Ranks. The formula for ρ can accommodate only two ranks; however, the data may consist of three or more sets of ranks. One method of summarizing the degree of agreement among three or more sets of rankings is simply to compute the mean of all possible ρ's. Thus, if three judges were to rank six pictures, we would compute ρ for every possible combination of paired sets in order to determine the average agreement among them. The paired sets would consist of the following combinations: Judges I and II, II and III, and I and III. The three ρ-values would then be averaged, *signs observed.* The result of this operation is sometimes called the *average intercorrelation of ranks;* however, it might be more properly called, for reasons given below, a *coefficient of agreement.*

Table 10.1.3 presents hypothetical rankings by three different judges. The average of the three ρ-values indicates only a moderate degree of agreement. Now let us suppose that we have three ρ-values of 1, −1, and −1. The three perfect correlations do not yield an average of 1.00 as one might naively expect, but only an average of −.33. This is comprehensible only if we view this average as a statement of degree

Table 10.1.3

Hypothetical Rankings by Three Judges

Judges' Rankings:			
I	II	III	
1	3	4	$\rho_{12} = .77$
2	2	6	$\rho_{23} = .26$
3	1	1	$\rho_{13} = \underline{-.09}$
4	4	2	.94
5	5	3	$\bar{\rho}$* = .31
6	6	5	

* Read "rho-bar."

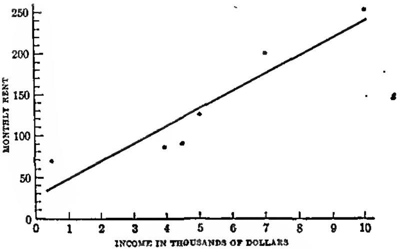


FIGURE 10.1.1 *Scatter Diagram, Monthly Rent and Family Income*

ready had occasion to note that rank correlation is not based strictly on a oorrespondence between rank orders, but rather on the correspondence between deviations of the respective ranks from the mean rank. These deviations are measured in terms of sigma units. The formula automatically transforms ranks into standard deviates and hence actual differences between ranks (D) into σ-differences. Therefore, we may eonfidently feed the original rank-order data into the machine-like formula.

However, the above set of operations proceeds on the assumption that intervals between ordinal ranks are equal. For instance, the gap between Ranks 1 and 2 is assumed to be equal to that between Ranks 2 and 3. And yet, our common sense tells us that the ideal circumstance of neat equal intervals which the formula requires can never be matched by the brute data which are ranked. The actual degree of subjective preference of first over second rank is not necessarily of the same intensity as third choice over second. Horses may rank in a certain order of passing the finish line, but the intervals between them are not the same. Nevertheless, ρ may be, and is, used when units of measure are unavailable, or when the discrimination between unequal intervals among variates is not considered essential.

From this it is evident that ρ measures the correlation between ordinal ranks, and not the correlation between the potential magnitudes that are being ranked. Hence, ρ in general overstates the degree of congruity existing between the raw variates, expressed or unexpressed. Thus, two

Table 10.1.4

Average Intercorrelation of Ranks

I	II	III	S	S^2
1	3	4	8	64
2	2	6	10	100
3	1	1	5	25
4	4	2	10	100
5	5	3	13	169
6	6	5	17	289
				747

$$\begin{aligned}\bar{\rho} &= 1 - \left[\frac{3(24+2)}{(2)(5)} - \frac{12(747)}{3(2)(35)(6)}\right]\\ &= 1 - \left[\frac{78}{10} - \frac{249}{35}\right]\\ &= 1 - [7.80 - 7.11]\\ &= .31\end{aligned}$$

Questions and Problems

1. Define the following concepts:
   Rank
   Rank Order
   Equal Intervals
   Ordinal Numbers
   Rank-Order Correlation

2. State the two fundamental difficulties in interpreting rank orders as indicators of the social realities which they purport to represent.

3. Under what circumstances can qualitative data be ranked?

4. Are differences between adjacent ranks necessarily equal? Illustrate.

5. Differentiate between agreement and correlation.

6. Verify by an example that tied ranks reduce rank-order correlation.

7. Two persons ranked six occupations by degree of prestige in the following orders. Compute ρ.

Minister	Physician
College Teacher	Banker
Banker	Lawyer
Lawyer	Engineer
Physician	Minister
Engineer	College Teacher

8. Four persons ranked the following items according to their conception of the degree of social distance registered by an endorsement. Compute the average intercorrelation. "I would be willing to have members of a given race:
   (a) live in my neighborhood."
   (b) live in this country."

of agreement, rather than of correlation. In this case of three perfect correlations, the predominant relation is one of disagreement, averaged with one of perfect agreement.

$$\begin{array}{lrr}
\text{Judge I and II} & \rho = & 1 \\
\text{Judge II and III} & = & -1 \\
\text{Judge I and III} & = & \underline{-1} \\
 & & \overline{-1} \\
 & \bar{\rho} = & -.33
\end{array}$$

Although the averaging of indexes of any type is usually fraught with pitfalls, it is not unjustifiable in this instance, for the reason that all the ρ-values are equally weighted, all being made up of the same number of items. It should be reiterated, however, that this average is not itself a coefficient of correlation, it is an average of several coefficients. It could never yield a minus unity, for the simple reason that if two series were correlated -1, a third could not be so correlated to both of the preceding. An average of plus unity, however, could be achieved in case of universal agreement.

When the ρ-values to be averaged are numerous, the computations will become laborious. But a shorter method of averaging ρ's has been devised,* which is particularly useful when the number of series is large. The formula is not as formidable as it appears at first view, since all of its terms are quite conventional quantities.

$$\bar{\rho} = 1 - \left[\frac{a(4N+2)}{(a-1)(N-1)} - \frac{12\Sigma S^2}{a(a-1)N(N^2-1)}\right] \qquad (10.1.2)$$

in which a = the number of sets
N = the number of items ranked
S = the sum of individual ranks of each item

Table 10.1.4 applies this formula to the data of Table 10.1.3 and the same $\bar{\rho}$-value is obtained (.31).

Utility of Rho. Since ρ applies to ordinal data, without specific units of measure, it fills an important need of the social scientist who must deal in such coarse quantitative measures, which are frequently of a subjective character. Sociometric rankings of preferences, aesthetic judgments, and the like may all be correlated in order to determine the degree of agreement among the ratings of selectors and judges. Occupations may be ranked by socio-economic level, and by their respective delinquency or divorce rates. Hence, a measure of presumptive relation between socio-economic level and incidence of divorce could thereby be obtained.

* Charles C. Peters and Walter R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases.* McGraw-Hill Book Co., New York, 1940, p. 201.

12. Calculate average intercorrelation by "short" metbod for Table 10.1.7 and interpret.

13. Baseball fans like to argue the question: "Which is more important for a team, offensive strength (i.e., tbe ability to hit and score runs) or defensive strength (i.e., good pitching and fielding, the ability to keep the other team from scoring runs)?" Table 10.1.8 shows the final standings of the major league teams in 1959, the number of runs scored by each team, and the number of runs scored against each team. Compute the rank-order correlation, for each league separately, between (a) final standing and runs scored and (b) final standing and runs yielded to the other teams. On the basis of these ρ-values, which factor seems to have been more important in the major leagues in 1959 — offense or defense?

Table **10.1.6**

Selected States Ranked for Efficiency in Elementary Education in 1860 and for Number of Persons per 1,000 Population Listed in Who's Who?, 1910

STATE	EDUCATION	WHO'S WHO?
Alabama	24	23
Arkansas...	29	27
Connecticut	2	2
Delaware. .. :.	8	19
Florida.........	27	29
Georgia.........	25	25
Illinois..........	14	14
Indiana.........	17	16
Iowa...	16	12
Kentucky.......	20	22
Louisiana.......	26	20
Maine..........	6	4
Maryland.......	13	15
Massachusetts...	1	1
Michigan.......	11	9
Mississippi	28	17
Missouri........	19	18
New Hampshire .	5	5
New Jersey......	9	13
New York.......	7	7
North Carolina..	22	28
Ohio............	10	11
Pennsylvania....	12	10
Rhode Island ...	4	8
South Carolina ..	21	21
Tennessee.......	23	24
Vermont........	3	3
Virginia.........	18	26
Wisconsin.......	15	6

Source: J. M. Reinhardt and G. R. Davies, *Principles and Methods of Sociology*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1932, p. 71.

(c) only as speaking acquaintances."
(d) as husband (wife)."
(e) as work companions in the same plant or store."
(f) as personal physician."

I	II	III	IV
a	d	b	a
b	b	c	c
c	c	d	b
d	a	f	d
e	f	e	e
f	e	a	f

9. Ten race-horses, designated A, B, C, D, E, F, G, H, I, J finish in the following order: H, C, D, F, G, A, J, B, E, I. Their position at the post was as follows: D, F, I, E, J, A, C, G, B, H. Correlate positions at the post with their finishing order.

10. Compute ρ between income and psychosis rate; occupational prestige and psychosis rate (Table 10.1.5).

11. Compute ρ for Table 10.1.6, and interpret.

Table 10.1.5 *Occupational Groups Ranked by Income, Social Prestige, and Psychosis Rate*

Occupational Group	Income	Prestige	Psychosis Rate
Peddlers	17	17	2
Waiters	16	16	1
Domestics	15	14	4
Barbers, beauticians	14	13	7
Semi-skilled and unskilled	13	15	3
Salesmen	12	9	8
Skilled workers	11	12	6
Office employees	10	8	14
Semi-professional (druggists, osteopaths)	9	4	9
Small tradesmen	8	5	15
Sub-executives	7	6	10
Policemen, firemen	6	10	13
Major salesmen	5	7	16
Minor government employees	4	11	5
Clergy, teachers, social workers	3	2	12
Technical engineers	2	3	11
Large owners, doctors, lawyers, dentists	1	1	17

Source: Robert C. Clark, "Psychoses, Income and Occupational Prestige," *American Journal of Sociology*, LIV, 1949, p. 433 Copyright by the University of Chicago.

SECTION TWO

Scatter Diagram and Correlation Table

In the previous section, we were concerned with the problem of measuring the amount of agreement between ranks, or ordinal measures. However, there are numerous concepts in sociology that are subject to interval measurement: for example, we may measure fertility in terms of the number of births per 100 women or wages in terms of dollars and cents. When the data being correlated consist of scaled variables, we naturally employ techniques that are appropriate to that type of quantitative data. While these correlational techniques differ somewhat from those previously presented, they all answer to the same purpose: namely, to express as precisely as possible the degree of relationship between two or more variables. The ensuing discussion is restricted to the most prevalent of all correlation measures: the product-moment correlation coefficient (r) and the correlation ratio (η).

The Scatter Diagram. The measurement of covariation can be approached in a preliminary manner by means of the *scatter diagram*, which is related to *bivariate* data in much the same way as the histogram is to univariate data. It reveals at a glance the entire disposition of items, thereby enabling us to arrive at a rough but useful estimate of the strength of the correlation before we actually measure it. This indispensable device has already been employed to depict the trend of a time series (p. 99) and the degree of correspondence between ranks (p. 274); however, the details of its construction have not yet been elaborated.

To illustrate more fully its construction and use, we take as our point of departure Table 10.2.1, which presents yearly income averages and suicide rates for the U.S. Central states, 1929–1950. To reduce the data of this table to a scatter diagram, we first draw horizontal and vertical axes, as in the construction of a histogram. Axes are drawn approximately equal in length, unless there is good reason to deviate from this convention. Also, as a matter of convention, the independent X-variable is plotted along the base line, and the dependent Y-variable along the vertical axis. The establishment of one variable as the independent, and the other as dependent, is, of course, not a statistical problem, but rather a matter of judgment and circumstance. The dependent variable may be construed as the effect of the independent variable as cause, or the outcome to be predicted from the predictor variable. In many instances, there may be no clear "causal" dependency at all, since both variables may be viewed as the consequence of an unidentified third factor.

Scales are next established on the axes in such a manner as to accom-

Table 10.1.7 *Ranking of Composers by Designated Groups*

BOSTON SYMPHONY ORCHESTRA	MUSICOLOGISTS (1944)	MUSICOLOGISTS (1951)
1. Beethoven	1. Bach	1. Beethoven
2. Brahms	2. Beethoven	2. Bach
3. Wagner	3. Mozart	3. Brahms
4. Mozart	4. Wagner	4. Haydn
5. R. Strauss	5. Haydn	5. Mozart
6. Bach	6. Brahms	6. Debussy
7. Sibelius	7. Handel	7. Schubert
8. Tchaikowsky	8. Schubert	8. Handel
9. Debussy	9. Debussy	9. Wagner
10. Haydn	10. Tchaikowsky	10. R. Strauss
11. Schubert	11. R. Strauss	11. Tchaikowsky
12. Handel	12. Sibelius	12. Sibelius

Source: Paul R. Farnsworth, *The Social Psychology of Music*, Holt, Rinehart and Winston, Inc., New York, 1958, p. 168 (adapted).

Table 10.1.8 *Final Standings of Major League Teams: Games Won, Runs Scored, and Runs Yielded, 1959*

TEAM	GAMES WON	RUNS SCORED	RUNS YIELDED
National League			
Los Angeles	88	705	670
Milwaukee	86	724	623
San Francisco	83	705	613
Pittsburgh	78	651	680
Chicago	74	673	688
Cincinnati	74	764	738
St. Louis	71	641	725
Philadelphia	64	599	725
American League			
Chicago	94	669	588
Cleveland	89	745	646
New York	79	687	647
Detroit	76	713	732
Boston	75	726	696
Baltimore	74	551	621
Kansas City	66	681	760
Washington	63	619	701

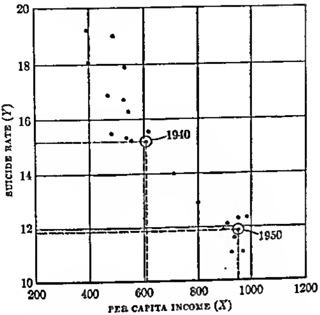


FIGURE 10.2.1 *Scatter Diagram, Suicide Rate by Annual Per Capita Income, Central U.S., 1929–1950*

tending perpendicularly from $Y = 15.2$ and $X = 602$; similarly, 1950 is represented by a point at the intersection of 11.9 and 947. The swarm of all such points constitutes the scatter diagram.

Types of Scatter. It is the pattern of this swarm that enables us to judge the nature of the relationship — and such a judgment is an essential preliminary to its proper measurement. Thus, it appears that a fixed increase in income is accompanied by a fixed decrease in the suicide rate; that is, the suicide rate changes by a constant amount per unit income. Such a relation is termed *rectilinear*, or simply *linear*, because the trend of scatter conforms to the track of a straight line.

Any such trend line, whether freehand or mathematically fitted, is technically termed a *line of regression*. This concept was coined by Galton in 1877, who used it in connection with his correlational studies of the characteristics of parents and their offspring. He perceived that such a line effectively expressed the tendency of children to "regress" to the average level of the parents in a wide variety of traits. The term has survived and enjoys a wide currency, although it is no longer restricted to its original connotation.

Table 10.2.1

Suicide Rate by Annual Per Capita Income, Central U.S., 1929–1950

YEAR	PER CAPITA INCOME (X)	SUICIDE RATE (Y)
1929	604	15.4
1930	532	17.9
1931	486	19.1
1932	399	19.3
1933	400	18.1
1934	437	16.9
1935	482	15.5
1936	549	15.2
1937	567	16.3
1938	515	16.7
1939	566	15.3
1940	602	15.2
1941	708	14.0
1942	800	12.8
1943	914	11.0
1944	972	11.0
1945	985	12.3
1946	951	12.3
1947	897	12.1
1948	928	11.6
1949	878	11.9
1950	947	11.9

Source Donald Faigle, *Suicide in Relation to Income, Urbanization and Race*, Unpublished Master's Thesis, Department of Sociology, Indiana University, 1956 Table 5

modate, with a margin to spare, the observed ranges of the respective variables. Thus, the horizontal scale covers the distance from $200 to $1,200, while the vertical scale extends from 10 to 20. *Needless to say*, enough markers are set up on each axis to ensure accurate and effortless plotting. Unlike the histogram, the vertical scale in a bivariate plot need not begin with zero, for the reason that the focus of attention is on the contour of the scatter rather than on relative frequency as gauged by the height of the curve. In this respect, the scatter diagram is analogous to the semi-log time chart, whose meaning likewise inheres in its shape rather than location.

Having drawn and scaled the axes, we are ready to plot each pair of values by a double-duty point. For any pair of variates, the Y-value fixes the height of the point above the base line and the X-value fixes its horizontal distance from the vertical axis. Thus, 1940 is represented in Figure 10.2.1 by a point located at the intersection of guide lines ex-

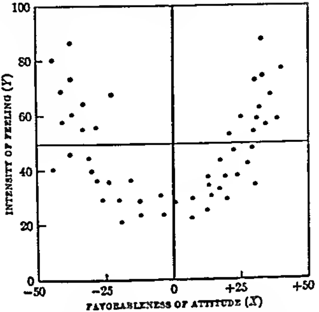


Figure 10.2.3 *Scatter Diagram, Intensity of Feeling by Favorableness of Attitude (Hypothetical Data)*

The three scatter diagrams presented thus far have distinctive trends and one would have little hesitancy in describing them as exclusively linear or curvilinear. But scatters of empirical observations are seldom so clean and unambiguous; more often both linear and curvilinear tendencies combine in the same set of data and thereby complicate the problem of representing correlation by an over-all measure. For example, the scatter of delinquency rates and average monthly rentals for 140 small census tracts in the city of Chicago (Figure 10.2.4) is marked by some linearity, yet it appears that a curved trend line would better fit the entire scatter. From left to right, as rentals increase, the delinquency rate responds by decreasing, but at a progressively slower rate. This is evidenced by the straightening of the swarm. There are of course several striking exceptions to the foregoing generalization — a few extremely high delinquency rates occur with above-average rentals. These mavericks would require special analysis, since they represent a breakdown in the "law." Yet, in the main, the law of relationship holds fairly well, affording a measure of predictability of one variable from the other. If we knew, for example, the rental to be $60, we would predict the delinquency rate to be approximately 3.0, which is the height of the free-hand *regression curve* at that point. To be sure, such a prediction would not

Furthermore, the relation in Figure 10.2.1 is said to be *inverse*, since the two series move in opposite directions: as income rises, the suicide rate falls. If suicide and income had risen together, then the trend of points would have been upward, reading from left to right. Such a sloping trend line (Figure 10.2.2) is evidence of a *direct* linear relationship between two variables; family size increases as farm acreage increases.

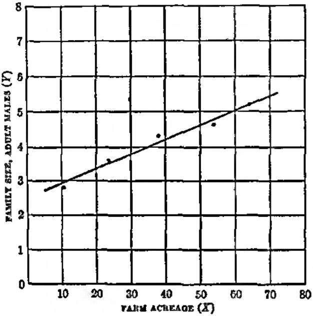
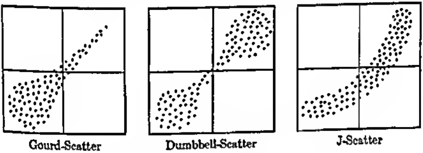


FIGURE 10.2.2 *Family Size by Farm Acreage*

The trend of the scatter will not always be linear; rather it may be *curvilinear* and take on any one of innumerable curve patterns. A simple example is provided in Figure 10.2.3, which portrays the relation between favorableness of attitude toward a minority people and intensity of feeling. As might be anticipated, a decided opinion — whether pro or con — is held with considerable intensity of feeling, while a less decided or neutral opinion arouses no very strong feelings. Whether this pattern — strong opinion, strong affect; weak opinion, weak affect — holds under all conditions is not here our concern. That would be a matter for empirical investigation. Here we are merely interested in exhibiting a *type* of relationship which is fairly common in sociological studies.

scattered for that value; on the other hand, for a rental of $70, we could predict with much greater accuracy, owing to the hunching of delinquency rates around the regression line. This scatter of Y-values for the respective X-values is known as ***scedasticity***. If the degree of variation in delinquency rates — the width of the scatter band — had heen uniform for all values of X, then we could have spoken of Y as being *homoscedastic* in respect to X. Actually, the degree of scatter in Y diminishes as X changes, so that Y is ***heteroscedastic*** in respect to X. Heteroscedasticity implies that the degree of correlation is not uniform throughout the entire series; hence, its presence reduces the feasibility of a single over-all measure of correlation, which is after all an average. Just as we hesitate to compute the mean of heterogeneous bimodal data, similarly we hesitate to calculate an average measure of correlation of a heteroscedastic scatter.

Heteroscedasticity may be even more glaring tban that exhibited in Figure 10.2.4; the scatter may be gourd-shaped, dumbbell-shaped, or J-shaped, as in Figure 10.2.5. These oddly shaped scatter diagrams by


Gourd-Scatter

Dumbbell-Scatter

J-Scatter

Figure 10.2.5 *Selected Types of Scatter*

no means exhaust the variety of types that may be encountered in practical work. However, they do serve to ratify the utility of this visual aid. Although the scatter diagram yields no mathematical measure of correlation, it does indicate (a) whether the relationship is simply rectilinear or more complex, (b) whether or not the relationship is consistent over the entire range, and (c) whether the relation is strong or weak. It is an indispensable tool and plays the same role in correlation as does the frequency graph in the processing of univariate data. It provides a hird's-eye view of the whole distribution.

The Joint Frequency Table.. Instead of plotting tbe individual items of the bivariate data, we may group them. Such grouping answers the general purposes of all grouping: (1) to reveal the basic pattern of dis-

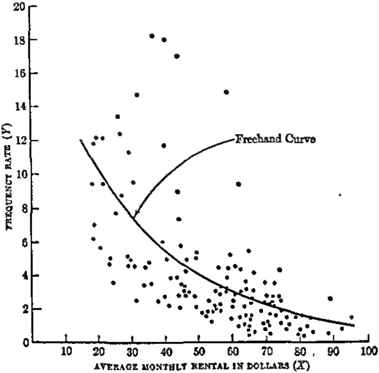
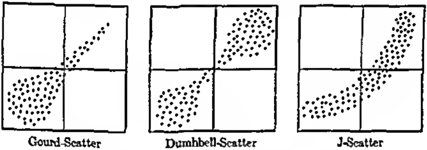


FIGURE 10.2.4 *Scatter Diagram and Freehand Regression Line, Delinquency Rate by Monthly Rental, 140 Local Areas, Chicago, 1930*

be free of error, for the obvious reason that none of the observed values fall right on the curve at that point — all deviate to a greater or lesser extent. Evidently, the accuracy of any such prediction would vary according to the tendency of the points to hug the line of relationship between the two series. When the points move within a narrow lane, predictive accuracy, and therefore correlation, would be high; when the points are widely scattered, predictive accuracy would be correspondingly low. Only when all points fall right on the regression line would prediction and correlation be perfect. At the other extreme, when the scatter is purely random, then we may just as well ignore the so-called "predictor variable." For any or all rentals, our best guess would be the over-all mean of the delinquency rates.

Scedasticity. Knowing the rental to be $30, we still could not accurately forecast the level of delinquency, since the delinquency values are widely scattered for that value; on the other hand, fnr a rental of $70, we could predict with much greater accuracy, owing to the bunching of delinquency rates around the regression line. This scatter of Y-values for the respective X-values is known as *scedasticity*. If the degree of variation in delinquency rates — *the width of the scatter* band — had *been uniform* for all values of X, then we cnuld have spoken of Y as heing *homoscedastic* in respect to X. Actually, the degree nf scatter in Y diminishes as X changes, so that Y is *heteroscedastic* in respect to X. Heteroscedasticity implies that the degree of correlation is not uniform throughout the entire series; hence, its presence reduces the feasihility nf a single over-all measure of correlation, which is after all an average. Just as we hesitate to compute the mean of heterogeneous bimodal data, similarly we hesitate to calculate an average measure of correlation of a heteroscedastic scatter.

Heteroscedasticity may he even more glaring than that exhibited in Figure 10.2.4; the scatter may be gourd-shaped, dumbbell-shaped, or J-shaped, as in Figure 10.2.5. These oddly shaped scatter diagrams by no means exhaust the variety of types that may be encountered in practical work. However, they do serve tn ratify the utility nf this visual aid. Although the scatter diagram yields nn mathematical measure of correlation, it does indicate (a) whether the relationship is simply rectilinear or more complex, (b) whether or nnt the relationship is consistent over the entire range, and (c) whether the relation is strong nr weak. It is an indispensahle tool and plays the same role in correlation as does the frequency graph in the processing of univariate data. It provides a hird's-eye view of the whole distribution.



Gourd-Scatter

Dumhbell-Scatter

J-Scatter

FIGURE 10.2.5 *Selected Types of Scatter*

The Joint Frequency Table.. Instead of plntting the individual items of the hivariate data, we may group them. Such grouping answers the general purposes of all grouping: (1) to reveal the basic pattern nf distribution, and (2) to facilitate the statistical processing of the data. In grouping bivariate data, we classify each case simultaneously in two class intervals, thereby locating each case at the intersection of a given row and given column. Hence, it is as though we superimposed a grid on the scatter diagram, counted the points in each cell and inserted the corresponding number. Such an operation, applied to the scatter of delinquency rates and average rentals (Figure 10.2.6), would yield the joint frequencies of Table 10.2.2.

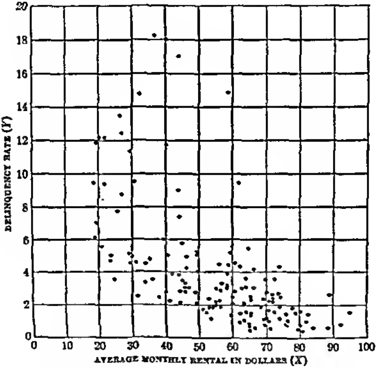


FIGURE 10.2.6 *Grid on Scatter Diagram, Delinquency Rate by Average Monthly Rental*

Naturally, in any real situation we would not proceed in this somewhat fanciful manner; rather, we would immediately tally the unarrayed bivariate items in the grid set up for that purpose, and then count the joint occurrences in each cell.

Table 10.2.2 *Joint Frequency Table; Delinquency Rates by Average Monthly Rentals*

Delinquency Rate	Average Monthly Rental in Dollars (X)										Frequency (f_y)
(Y)	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–99	
18.0–19.9				1	1						2
16.0–17.9					1						1
14.0–15.9				1		1					2
12.0–13.9			4								4
10.0–11.9		1	1		1						3
8.0– 9.9		1	2	1	1		1				6
6.0– 7.9		2	1	1	1						5
4.0– 5.9			5	4	6	3	4	1			23
2.0– 3.9			1	4	14	11	12	9	1		52
0– 1.9						7	13	13	7	2	42
Frequency (f_x)	0	4	14	12	25	22	30	23	8	2	140

Grouping Procedure. To clarify this grouping procedure, we list the directives observed in the construction of Table 10.2.2.

1. Select a suitable class interval for each variable.

2. Mark off these class intervals on properly drawn axes, and from these markers extend guide lines in such fashion as to create a gridwork of cells. Each cell is taken to represent the junction of two class intervals.

3. Locate each pair of values in its proper cell and indicate its presence there by a tally mark.

4. Count tally marks in each cell and replace tally by that number. Any such number represents a joint frequency — i.e., the frequency with which two class values or midpoints, occur together.

5. Sum by rows and columns in order to obtain the marginal totals. These marginal totals of course constitute the simple frequency distributions of the individual variables.

6. Sum marginal frequencies to obtain the grand total of cases, N, the one sum serving as a check on the other.

In retrospect, we see that the construction of the joint frequency table proceeds according to well-established principles: the classification should be fine enough to reveal the shape of the joint frequency distribution, but not so fine that some rows and columns are completely vacant. But if one is to err at the outset, it is perhaps preferable to have too many cells than too few. The optimum arrangement will be more apparent in the light of too much detail rather than too little, and it is usually easier to amalgamate cell frequencies than to partition them.

Function of the Joint Frequency Table. We must acknowledge that a joint frequency table cannot depict the type of relationship so vividly as the *scattergram*, for the reason that numerals cannot convey so effectively gradations of density as does the scatter of dots. Nevertheless, in spite of its relative coarseness, it will often be more than adequate for purposes of exhibiting the pattern of relationship and assuring its comprehension.

Moreover, the column and row distributions of the joint frequency table permit a statistical measurement of scedasticity which would be impossible to obtain from a raw scatter diagram. For such an assessment, *we need only compute the standard deviation of each column or row.* A close similarity among such standard deviations would be evidence of homoscedasticity, whereas pronounced differences would suggest heteroscedasticity.

Furthermore, the marginal distributions of the X- and Y-variables — *which, of course, are integral features of the joint frequency table* — also supply important and recognizable clues to the possible degree of association between the plotted variables. Specifically, the marginal distributions set limits to the degree of obtainable correlation. For example, unlike marginals preclude a perfect linear relationship. This merely generalizes the earlier observation in respect to the ϕ-coefficient that perfect two-way association in contingency tables is impossible so long as marginal sets are not identical. This general principle is illustrated by Table 10.2.3, whose marginals make it impossible to locate

Table 10.2.3

Unlike Marginal Sets

				f_y
				10
Y-Variable				10
				10
f_x	15	10	5	30
		X-Variable		

all cases along one diagonal, as would be required in perfect rectilinear correlation. No matter how we deploy the eases, we cannot force them onto one diagonal, as the student should experimentally demonstrate for himself. In general, the marginals always act to constrain in one way or another the type and degree of association between paired variables; hence, it is always necessary to examine them in order to determine the limits which they impose, and the prospects for the valid employment of a given correlation index.

Additionally, the joint frequeney table may serve as a computing aid which may be called on especially when the number of instances is large or when computing machines are not available. Finally, it is a prerequisite to the simplest approach to curvilinear correlation, as will be evident in a later section.

This section has stressed the important role of the scatter diagram and the joint frequency table in the preliminary analysis of covariation. From these charts it is possible to determine whether the relationship is (a) linear or curvilinear; (b) direct or inverse; (c) weak or strong; additionally, (d) whether the variables are homoscedastic in respect to each other; (e) whether there are notable exceptions to the main trend or "law of relation," and (f) whether the marginal distributions are symmetrical or skewed, and approximately matched. Because of these significant disclosures, it is mandatory that, prior to the measurement of correlation, a scatter diagram or joint frequency table, or even both, be constructed and carefully studied. Without these visual aids, we fall into the danger of using improper procedures and thereby arriving at misleading conclusions.

Questions and Problems

1. Define the following concepts:
   Scatter Diagram
   Bivariate Data
   Rectilinearity
   Scedasticity
   Curvilinearity
   Joint Frequency Table
   Marginal Distributions

2. If the marginal distributions are normal in shape, will the scatter necessarily be homoscedastic? Illustrate and explain.

3. Is a perfect rectilinear scatter possible, if marginal distributions are unlike? Illustrate your answer by table or graph.

4. Does the presence of homoscedasticity in a scatter guarantee linearity? Illustrate by sketch.

5. Does perfect rectilinearity in a scatter require identical marginals? Illustrate.

Table 10.2.4

Clothing Expenditures by Total Family Expenditures, U.S.

TOTAL EXPENDITURES ('00 OF DOLLARS) (X)	CLOTHING EXPENDITURES (PER CENT OF TOTAL) (Y)
6	*5*
10	8
14	12
18	13
20	14
24	15
28	16
34	18
43	20

Table 10.2.5

Per Cent Democratic Vote by Voter Registration, Selected States, Congressional Elections, 1952

STATE	DEMOCRATIC VOTE (% OF TOTAL)	PER CENT REGISTERED
Ariz.	52	78
Ark.	85	50
Cal.	44	88
Colo.	44	91
Conn.	46	91
Fla.	*74*	*75*
Ga.	100	59
Idaho	41	96
Ind.	43	96
La.	91	67
Md.	48	62
Mass.	46	88
Mont.	43	83
Nev.	50	81
N.H.	37	92
N.J.	42	85
N.Y.	40	80
Ohio	44	59
Ore.	39	87
Pa.	48	78
R.I.	54	87
S.C.	93	49
S.D.	31	83
Vt.	28	88
Va.	67	33
Wash.	43	92

Source: U.S. Bureau of the Census, *Statistical Abstract of the U.S., 1952*, U.S. Government Printing Office, Washington, D.C., 1952

6. If a scatter is homoscedastic, does it follow that the relationship is close?

7. (a) Construct a scatter diagram for the data in Table 10.2.4.
   (b) Formulate in your own words the "law of relation."

8. (a) Plot a scatter diagram for the following data of Table 10.2.5.
   (b) Characterize the relationship. Does a preponderant one-party vote appear to depress voter registration?

9. Construct the scatter diagram for the delinquency and recidivism rates (Table 10.2.6). How would this trend be described in words?

Table 10.2.6

Delinquency and Recidivism Rates, Chicago, 1929

Delinquency Rate per 100	Recidivism Rate
0.5	26.1
1.5	26.6
2.5	35.1
3.5	30.1
4.5	38.7
5.5	54.8
6.5	42.9
7.5	54.7
8.5	32.4
9.5	52.3
10.5	66.7
11.5	50.2
12.5	59.9
13.5	65.9
15.5	53.6
16.5	80.3
19.5	61.1

Source: Clifford R. Shaw, *Delinquency Areas*, The University of Chicago Press, 1929, Table 21, p. 181.

10. (a) Plot a scatter diagram from Table 10.2.7.
    (b) What does the scatter of this graph suggest concerning the "pull" of economic conditions? Explain the two fairly distinct clusters.

11. Construct a joint frequency table for the data in Table 10.2.7. Divide Income Index into ten class intervals of width 10, lower rounded limit 50; divide Per Cent Born in Other States into ten class intervals of width 10, lower limit 0. Tally marginal frequencies.

12. Sketch the marginal distributions which would yield the scatters shown in Figure 10.2.7.

Table 10.2.7 *Per Cent Born in Other States by Income Index, U.S.*

STATE	INDEX OF INCOME *	% BORN IN OTHER STATES	STATE	INDEX OF INCOME	% BORN IN OTHER STATES
Ala.	62	11	Neb.	96	25
Ariz.	91	44	Nev.	137	68
Ark.	58	22	N.H.	99	30
Cal.	124	58	N.J.	120	32
Col.	99	50	N. Mex.	81	44
Conn.	127	28	N.Y.	124	17
Del.	138	35	N.C.	64	12
Fla.	80	54	N. Dak.	75	23
Ga.	69	14	Ohio	115	24
Idaho	88	49	Okla.	78	39
Ill.	121	25	Ore.	106	57
Ind.	103	25	Penn.	104	12
Iowa	94	19	R.I.	101	24
Kansas	104	32	S.C.	67	11
Ky.	69	12	S. Dak.	77	29
La.	74	15	Tenn.	69	20
Me.	83	12	Texas	89	21
Md.	107	34	Utah	88	21
Mass.	107	15	Vt.	82	22
Mich.	111	28	Va.	81	25
Minn.	91	21	Wash.	110	54
Miss.	50	11	W. Va.	75	17
Mo.	97	25	Wis.	101	15
Mont.	104	44	Wyoming	98	60

* Ratio of state per capita income to national per capita income.

Source: U.S. Bureau of the Census, *Statistical Abstract of the U.S., 1954*, U.S. Government Printing Office, Washington, D C , 1954, Tables 37 and 337.

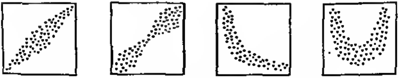


FIGURE 10.2.7 *Selected Scatter Patterns*

SECTION THREE

Linear Correlation of Two Variables

The Need for an Over-All Measure of Correlation. The scatter diagram, which was elaborated in the preceding section, reveals to the eye whether two variables change together in a systematic manner. The cluster of plotted points around the hypothetical regression line suggests the "law of relation" between the two variables. Moreover, by observing the width of the scatter, we can make at least a preliminary judgment as to how well the cases conform to this hypothetical law. Such conclusions are often of considerable value; yet they still leave something to be desired, for they are impressionistic, subjective, and unstandardized. Consequently, they cannot be described or communicated to anyone else without reproducing the chart — or at least engaging in an extended account of it. This is, of course, impractical. And since these "eye-measures" are not accurate in a mathematical sense, comparisons with similar measures are impossible even if all the intricate charts are available. What we need therefore is an objective, standard, synoptic *measure* of the relation between the two variables. This will finally be the measure of correlation.

In principle, there is of course no difference between rough measures made by the eye and those computed by a mathematical operation. The latter are merely more precise and serviceable. What we see is a hypothetical trend line to which the swarm of points more or less conforms and from which we deduce the degree of correlation. But in order to measure this correlation we must (1) establish exactly the position of such a line, (2) measure the amount of conformity thereto, and (3) translate this result by a prescribed method into an index of correlation. Our first concern, then, would be with the location of the line which best fits the observed data.

As we have previously learned, a relation may be rectilinear or curvilinear; hence, the best-fitting line may be straight or curved. In this section, we restrict ourselves to those relations which are rectilinear, or approximately so, and which can therefore be legitimately represented by a straight line. In other words, our concern will be with those relations in which a constant change in one variable produces, on the average, a constant change in the other paired variable. Such a relation has already been illustrated in Figure 10.2.2, which plots the size of family and the size of farm in China: family size increases by one person on the average as farm size increases by ten acres.

Here, for purposes of exposition, let us use some very simple data which exemplify the relation between income and social status score

(Table 10.3.1a). It is reasonable to designate social status as the dependent Y-variable, since status appears to adjust itself to income in the long run, rather than the reverse. Once cases have been plotted to form a scatter diagram (Figure 10.3.1), we wish to draw the most representative straight line.

Table **10.3.1a**

Income and Social Status Score, Selected Families

Income in Thousands of Dollars (X)	Social Status Score (Y)
3	3
7	5
11	7
14	6
15	9
$\bar{X} = 10$	$\bar{Y} = 6$

Source: Suggested by W. Lloyd Warner's studies in social class.

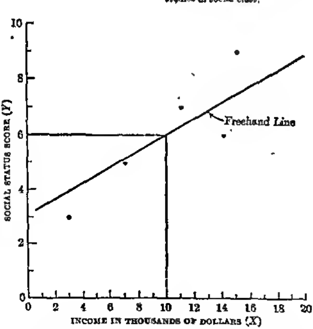


Figure 10.3.1 *Freehand Trend Line on Scatter Diagram, Income and Social Status*

Freehand Regression Line. It would of course be possible to draw freehand what appears by inspection to be the best-fitting trend line. If we had drawn such a line and were then asked to defend its location, we would spontaneously reply that we had drawn it "right through the middle" — that is, as close to all the points on the average as possible, cutting the swarm lengthwise into two equal bands. It was drawn down the middle because we intuitively felt that this line was the best average for *all* of the observed points. It thus serves as a running estimate of the Y-values, which takes into account every observed value of X. If we had been asked to guess a person's social status score from his income, say $10,000, we would not guess a Y-value outside, or even close to the edge, of the region of scatter, but rather a central value within the swarm above the designated X-value. This would be a status score of 6. Such *estimated values of* Y correspond to the observed X's and constitute the aforementioned trend line. Naturally, the more closely the observed points hug the line of estimation — the less they digress from it — the smaller the error in prediction of Y from X. If income were the sole determinant of social status, the observed points would necessarily all be on the line, without error; if social status alone determined income, the points would again fall on the line without residue.

Now, what is the significance for correlation of these departures from the hypothetical regression line? To the statistical analyst it indicates that there are *other* factors besides income which determine social status. The operation of these other, unknown factors disturbs our prediction. Hence, they reduce the correlation between the two given variables.

Values Expressed as Deviations from Mean. In Table 10.3.1a and the scatter diagram derived from it, the variates are given in their original form. But for several reasons, not fully amplified here, it is more meaningful to compare the variables after they have been transformed into deviations from their respective means, $\bar{X}$ and $\bar{Y}$.

When comparing values, as we do in correlation, the student must accustom himself to thinking in terms of deviations from the mean instead of the raw data. Like so many other statistical procedures which may at first seem needlessly indirect and complex, this too conforms to common sense. We do not ordinarily compare raw test scores, salaries, hirth rates; hut rather consider them as "high" or "low" — that is, above or below the average or norm to which we are accustomed. Still less can we compare units of different categories, such as incomes and hirth rates in their raw form; we associate income below average with family size that exceeds the average. Thus, we see that the mean is the appropriate point of origin — the natural standard — whenever we compare two sets of data in respect to their affinity with each other.

The items in Tahle 10.3.1h are therefore expressed as deviations from their respective means, and then plotted in that form (Figure 10.3.2a). In graphing deviations, we must of course be able to accommodate negative values, and consequently we use four quadrants, corresponding to the four combinations of signs.

Table 10.3.1b

Incomes and Social Status Scores as Deviations from $\bar{X}$ *and* $\bar{Y}$

INCOME IN THOUSANDS OF DOLLARS (x)	SOCIAL STATUS SCORE (y)
−7	−3
−3	−1
1	1
4	0
5	3
0	0

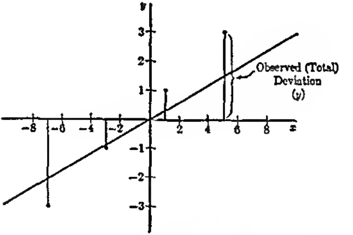
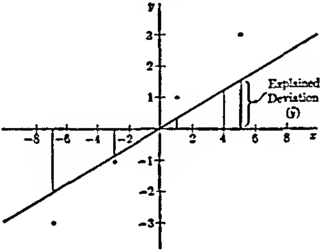


FIGURE 10.3.2a *Scatter Diagram, Observed Deviations* (y)

It should be observed that the configuration of plotted deviations is identical with the scatter of original values; only the scale markers have been changed. All we did was to shift the zero origin to the intersection of the means.

Explained Deviations. The freehand trend line, drawn straight through the middle, is obviously an average for all of the observed points, a continuous succession of norms, as it were, adjusted to the entire series of cases. Since it consists of the deviations which one would expect on the average, the values on the sloping trend line are variously called the "expected," "predicted," or "estimated" values. The expected deviations are graphically plotted (Figure 10.3.2b) by vertical lines extending



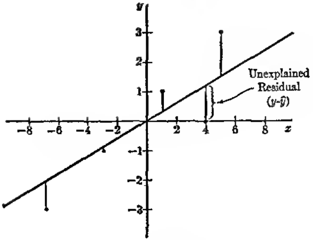


FIGURE 10.3.2c *Unexplained Residuals* $(y - \hat{y})$

explained; that is, they must be attributed not to income, but to unknown factors on the identity of which we may only speculate, since additional data are not available in the problem as set up. Hence, they are called *unexplained residuals*, measured by the respective distances from the trend line to the individual observed points in the swarm.

From inspection of the foregoing charts, we must conclude that the closer the correspondence between the explained and observed deviations (i.e., the smaller the residuals) the higher the degree of correlation. It would therefore be logical to measure the degree of correlation according to the degree of correspondence between the estimated and observed deviations; or by the proportion of the aggregate observed deviation that is explained. This is the fundamental principle of the measurement of correlation and is always implicit in its calculation. To be sure, the measurement is not carried out informally by freehand regression lines or by graphic readings obtained with a ruler or a pair of dividers; still, it remains essentially a comparison, or ratio, between the explained and total deviations.

Since these concepts play such an important part in the subsequent presentation, it will be well to summarize their meaning precisely:

(1) *Total* (*Observed*) *Deviation:* the deviation of an observed value from the mean of the series.

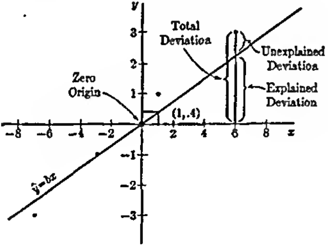
(2) *Explained Deviation:* the deviation of an expected (regression) value from the mean of the series.

(3) *Unexplained Residual:* the discrepancy between the total and the explained deviation. This is, of course, the difference between (1) and (2) above.

The Measurement of Linear Correlation

A serious disadvantage of freehand graphs is that they rest on personal judgment. Without a standard operating procedure, it is unlikely that two workers would ever locate the trend line in exactly the same position. Obviously, a line that is used to measure correlation in the manner suggested above must always meet the same specifications; otherwise the results will be lacking in the reliability which is essential to scientific procedure.

Such standard specifications have been formulated as follows: the line is so located as to make the sum of the vertical residuals around that line equal to zero — that is, to make the sum of the positive and negative discrepancies balance one another. Thus, such a line represents the scatter around it, as the mean represents its array. And like the mean, it makes the sum of the squared deviations a minimum — it conforms to the *principle of least squares.** Because of these properties it is labeled



the line of *best fit;* by the criterion of least squares, it fits the scatter better than does any other straight line.

The tracing of this mathematical line of best fit is simplified by again plotting cases as deviations from means. For, under these conditions, the line of least squares will always conveniently pass through the zero origin, *which lies at the intersection of the means. Consequently, it* may be readily plotted once its *slope* on the x-axis has been determined.

Computation of Slope. The calculation of this slope is according to the following equation, presented here without explanation:

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} \tag{10.3.1}$$

where b_{yx} = slope of line of expected y's on observed x's (read: "b, y on x")

Σxy = sum of products of paired x- and y-deviations (read: "sum of cross-products")

This ratio is the slope of the sought-for best-fitting line; it is the average change in Y per unit change in X. Accordingly, to find this average change in social status per unit change in income, we compute the cross-products and squared x-deviations, and form the ratio between their respective sums. Performing these operations (Tahle 10.3.2), we ohtain

Table 10.3.2 *Calculation of* b_{yx}

X	Y	x	y	x^2	xy
3	3	−7	−3	49	21
7	5	−3	−1	9	3
11	7	1	1	1	1
14	6	4	0	16	0
15	9	5	3	25	15
Σ = 50	30	0	0	100	40

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{40}{100} = .4$$

b_{yx} = .4. This is the average rate of increase in social status per unit of income; that is, for every $1,000 increase in income, there is an average increase of .4 status points. Because this average fixes the inclination of the regression line to the axis of the independent variable, it has come to be known as *slope.* Once determined, it enables us to plot the line of least squares and proceed to the precise measurement of correlation.

Plotting the Regression Line. Since a straight line is determined by any two points, we plot two points known to lie on the least-squares line and run a straight line through them. For one of these, we pick up the point of origin (intersection of axes); the other point is most conveniently obtained by calculating the expected value of y when $x = 1.00$. This will of course be equal to b_{yx}. Accordingly, we draw a line through the zero origin and a point plotted at the intersection of 1.00 and .4. This line is plotted in Figure 10.3.3.

It is this line that supplies the explained deviations ($\hat{y}$'s) corresponding to the observed x's, since the unbroken line of best fit necessarily consists of all possible explained deviations. From it, we may read the approximate magnitudes of the explained deviations; but they may be more precisely determined by application of the slope formula:

$$\hat{y} = b_{yx}(x) \tag{10.3.2}$$

where $\hat{y}$ = expected deviation in Y for any given x (read "y-circumflex"). Substituting the observed x-deviations in this formula, and solving for the $\hat{y}$'s, we obtain the results shown in Table 10.3.3. These $\hat{y}$'s are the y's that would have been observed if X and Y were perfectly correlated. But these $\hat{y}$'s were not observed; hence, we must determine how close they come to those observed, in order to fix the degree of correlation.

Table 10.3.3

Complete Set of $\hat{y}$-Values for Observed x's

b_{yx}	x	$\hat{y}$
.4	−7	−2.8
.4	−3	−1.2
.4	1	.4
.4	4	1.6
.4	5	2.0
	0	0.0

Coefficient of Determination. We have thus come to the final stage in the measurement of correlation. This measurement proceeds on the principle that the more nearly the explained deviations approach the total deviations, *obviously the greater is the proportion of variation explained and* the higher the degree of correlation. As a statistical operation, the conversion to a single index consists of summing the squared explained deviations, and expressing this *explained variation* as a proportion of the *total variation* to be explained. Symbolically,

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\Sigma \hat{y}^2}{\Sigma y^2} \tag{10.3.3}$$

This ratio, be it noted, is symbolized by r^2. It is the fraction of the total *variation in the dependent Y-variable* which is determined by the independent X-variable; hence, it is called the *Coefficient of Determination*. Since it is impossible to explain more than the total variation, r^2 can never be greater than unity, and will practically always be less than one.

Had the simple algebraic deviations around the mean been employed *as measures of variation, they would have summed to zero* — a very impractical situation which would not have permitted any ratios at all. Hence we turn to squared deviations, in compliance with the principle of least squares.

But in order to circumvent this more complicated computation, would *it not have been possible to take the simple arithmetic deviations from* the median? This was, in principle, the solution of Francis Galton who originally propagated the concepts of regression and correlation in his famous work *Natural Inheritance* (1889). However, the mean has properties not possessed by the median which make it more useful in correlation computations. And *once the mean has been chosen*, there is no alternative but to adopt the least-squares principle for its implementation.

The complete computation of r^2 is illustrated in Table 10.3.4, where the explained and observed deviations are squared and summed to give the explained and total variation, respectively. Upon dividing the total variation, 20, into the explained variation, 16, we obtain $r^2 = .80$. This is the final measure of the degree of association between the two variables: it reveals that 80 per cent of the variation in Y is accounted for by its linear dependence on X.

Table 10.3.4 *Computation of r^2*

x	y	b	x	$\hat{y}$	$\hat{y}^2$	y^2
−7	−3	(.4)	(−7)	−2.8	7.84	9
−3	−1	(.4)	(−3)	−1.2	1.44	1
1	1	(.4)	(1)	.4	.16	1
4	0	(.4)	(4)	1.6	2.56	0
5	3	(.4)	(5)	2.0	4.00	9
				0.0	16.00	20
$r^2 = \frac{16}{20} = .80$						

Reversibility of r^2. We could have computed the regression of X on Y as easily as the regression of Y on X, and thereby determined the proportion of X-variation explained by Y. In fact, such a result would appear essential to a complete statement of correlation. Why compute r^2_{yx} and ignore r^2_{xy}? The answer is that it has not been ignored, although we did not compute it independently. It is simply unnecessary to fit both regression lines for the reason that $r^2_{yx} = r^2_{xy}$. In short, r^2 is reversible. For example, knowing that $r^2_{yx} = .80$, we may state not only that income explains 80 per cent of the variation in social status but the reverse as well, namely, that social status accounts for 80 per cent of the variation in income. Since, from a purely statistical standpoint, each explains the other to the same degree, the subscript is usually not attached to r^2.

Coefficient of Non-determination. Since the difference between total variation and explained variation is necessarily equal to the unexplained variation, it follows that the unexplained proportion is simply the difference between unity and r^2, a quantity appropriately termed the *Coefficient of Non-determination*. We may write it as follows:

$$1 - r^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

We could have obtained this quantity directly by measuring the residuals around the regression line, *squaring* and *summing* them, and expressing this sum as a proportion of total variation. In fact, this is the operational meaning of $1 - r^2$. But whether calculated directly from the residuals, or indirectly via r^2, this coefficient may always be construed as the proportion of variation in the dependent variable left linearly unexplained by the independent variable. It thereby gauges the strength of the unidentified factors.

Pearsonian Product-Moment r. Conventionally, it is r rather than r^2 that is selected and quoted to indicate the degree of correlation between the two variables. Why is this so, particularly since r^2 seems to have all of the prerequisites for a satisfactory index of correlation: it ranges between zero and unity; it has a simple, comprehensible meaning? Nevertheless, unless another index is specifically named, it is assumed that *the* correlation coefficient refers to *Pearsonian r*. How is this convention explained?

Before formulating an answer to that question, we must first acquaint ourselves with the distinctive properties of r. Accordingly, we return once again to the scatter diagram and the line of best fit. Our procedure will be exactly the same as that followed previously, excepting that the data will now be plotted as standard deviates (σ-values), instead of the observed deviations (Figure 10.3.4). Again the configuration of points is left unaffected by the transformation: the sole alteration is that the scale unit is now the standard deviation rather than the original raw unit of measure.

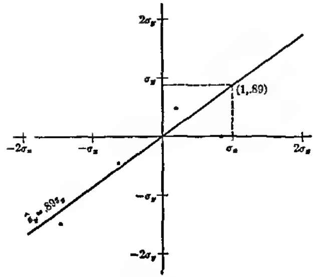


FIGURE 10.3.4 *Scatter Diagram of Standard Deviates*

Table 10.3.5

Deviations as Standard Deviates

DEVIATIONS		STANDARD DEVIATES	
x	y	$\frac{x}{\sigma}$	$\frac{y}{\sigma}$
−7	−3	−1.56	−1.50
−3	−1	− .67	− .50
1	1	.22	.50
4	0	.89	.00
5	3	1.12	1.50
0	0	0.00	0.00
$\sigma_x = 4.47$ $\sigma_y = 2.00$			

To establish the line of best fit through this scatter of σ-points, we first calculate its slope as before, except that now we operate on the standard deviates instead of the raw deviations. Consequently, the slope would now be written in standard form:

$$\frac{\sum\left(\frac{x}{\sigma}\right)\left(\frac{y}{\sigma}\right)}{\sum\left(\frac{x}{\sigma}\right)^2}$$

instead of:

$$\frac{\Sigma xy}{\Sigma x^2}$$

Since $\frac{x}{\sigma}$ is usually symbolized z, the slope formula may be simply written as:

$$r_{yx} = \frac{\Sigma z_y z_x}{N} \tag{10.3.4}$$

and symbolized r_{yx} instead of b_{yx}. This is the culmination of our development. We are now in a position to see that the slope of the regression line through the scatter of σ-points is identical with the square root of explained variation: that is, $r = \sqrt{r^2}$. For the example above (Table 10.3.4), $\sqrt{.80} = \pm .89$.

To compute r directly, we would proceed as in Table *10.3.6*, where paired standard deviates are multiplied together and those cross-products are summed and averaged. The mean of the cross-products is r.

Table 10.3.6

Pearsonian r as Mean of Cross-Products of Standard Deviates

z_x	z_y	$z_y z_x$
−1.56	−1.50	2.340
− .67	− .50	.335
.22	.50	.110
.89	.00	.000
1.12	1.50	1.650
		4.465

$$r = \frac{\Sigma z_y z_x}{N} = \frac{4.46}{5} = .89$$

We can now understand why r was designated by Karl Pearson "product-moment correlation coefficient." From the composition of Formula 10.3.4 above, it is clear that r is strictly an arithmetic mean; it is a sum divided by the number of items in that sum. Now, the terms "mean" and "moment" are used interchangeably in mathematical statistics; hence, Pearson labeled the mean of the standard cross-products the product-moment correlation coefficient.

Viewed in this light, r may be construed as the mean change in Y for every unit change in X, or the reverse, always assuming measures in standard form, i.e., expressed in terms of sigma units. Thus, if a given income deviates by 1σ from its mean, we expect the associated status scores to deviate on the average from their mean by $.89\sigma$. By virtue of this principle, we may estimate Y (in standard form) for any given dependent, r^2 measures the over-all proportion of the total variation of one variable that is associated with, or explained by, the other.

On the other hand, r measures the dynamic aspect of this relation, measuring the rate of change in one variable relative to the other, as has been demonstrated in the paradigms above. Because of this conceptual distinction, we may say that r is primarily a predictive device to forecast, for example, the expected level of performance on one variable from observed performance on another. As such, it is probably more likely to be used by educators, psychologists, and others interested in personal prediction, rather than by sociologists. On the other hand, r^2 is a summarizing measure weighing the influence, or force, exerted by one variable on the other.

Since r is slope, it manifestly must have a direction: predominantly up or down, according to whether the variables are positively or negatively related. It follows that the direction of the slope reflects the type of relation, which is then symbolized by a plus or minus sign. The synoptic measure r^2, however, carries no sign since it expresses a proportion of a total variation.

It is now clear that r and r^2 are not interchangeable; nor should they be cheaply derived from each other until their structural meanings are thoroughly understood. Since r is always larger than r^2, it could be deliberately or unwittingly used to exaggerate the strength of association and thereby mislead the reader: for example, $r = .5$ may seem to signify a reasonably strong association, but $r^2 = .25$ indicates that only 25 per cent of the variation in either variable is accounted for by the other. When the emphasis is on the strength of the over-all relationship between two variables, as is frequently the case in sociological studies, r^2 is the pertinent statistic. For example, the finding that 50 per cent of the variation in area delinquency may be attributed to the economic factor still leaves something unexplained; yet it represents a tangible gain in understanding a phenomenon that until a century ago was laid at the door of demons, heredity, or free will.

In sum, each measure has its very distinctive connotations and appropriate uses. Nevertheless, r enjoys a near monopoly over r^2. This may be due partly to the inertia of tradition, which will perhaps be dissipated when workers become more sensitive to the nuances of quantitative reasoning.

Computing Formulas: Ungrouped Data. The extended analytic operations previously given were designed to expose the logic and intent of the concepts of r and r^2, so that they could be discriminatingly applied in social research. But such elaborate calculations need not be carried out if the sole object is to determine the mere numerical value of r or r^2 in a given practical problem. There are handbook working procedures value of X, exactly as we applied b_{yx} to obtain the expected y-deviations. We have only to apply the formula:

$$\hat{z}_y = r(z_x) \tag{10.3.5}$$

where $\hat{z}_y$ = estimated σ-value in Y
z_x = observed σ-value io X

Substituting our data in this formula, we get the estimated values shown in Table 10.3.7. Thus, when iocome is 1.12 in etandard form, social status is estimated to be 1.00; when iocome is -1.56σ, social status would be -1.39σ. Perfect correlatioo between two eeries would prevail whenever the paired values are identical sigma distances from their respective means — for example, if the person with a sociology grade of, say, 1.8σ above tbe mean were located 1.8σ above the mean of the history grades, and all other observatione in the two series were also perfectly paired. Because r is reversible, we could have applied r to the observed z_y's and thereby estimated the corresponding z_x's.

Table 10.3.7 *Estimated z_y from Observed z_x*

z_x	$(r)z_x$	=	$\hat{z}_y$
−1.56	(.89)(−1.56)	=	−1.39
− .67	(.89)(− .67)	=	− .60
.22	(.89)(.22)	=	.20
.89	(.89)(.89)	=	.79
1.12	(.89)(1.12)	=	1.00

Comparison of r and r^2. On the surface, the difference between r and r^2 appears very trivial: a simple detail of exponent, with both values easily convertible into the other, after one has been computed. Nevertheless, this seemingly innocent difference caooot be so lightly dismissed, for the two r's respectively focus on two distinct but interrelated aspects of covariation — a distinction which the design of the present chapter has deliberately sought to portray.

Either one could have been mechanically derived from the other by way of only one of the two procedures. However, an r^2 derived from r would never convey to the student the intent and meaning of explained variation which r^2 measures. The "square of the slope" would mean nothing to him. Nor would an r, converted arithmetically from r^2, convey the intent and meaning of the slope. The "square root of explained variation" could not be visualized as slope Although r and r^2 are inter-

Grouped Data. The computing procedure for grouped data is essentially the same as for ungrouped, except that we operate on coded class midpoints instead of the unarrayed, individual values:

$$r = \frac{N\Sigma fx'y' - \Sigma fx'\Sigma fy'}{\sqrt{[N\Sigma fx'^2 - (\Sigma fx')^2][N\Sigma fy'^2 - (\Sigma fy')^2]}} \qquad (10.3.8)$$

To illustrate Formula 10.3.8, we apply it to the correlation table of delinquency rates and average rentals (Table 10.2.2, p. 289), reviewing at the same time the construction of that table. The steps in the procedure are as follows:

1. Prepare a grid having at least as many rows and columns as class intervals in X and Y. Our grid has 10 rows and 10 columns.

2. Write down the X-intervals of uniform width in the upper margin as column heads; write down the Y-intervals in the left-hand margin as row heads.

3. Turning to the raw, unordered data, cross-tabulate each item by placing a tally mark in the appropriate cell. Count the tally marks in each cell and recoed the number in the center of the cell.

4. Sum the row frequencies, recoeding them in the column labeled f (Table 10.3.9); sum the column frequencies and record them in the row labeled f. Then sum these row and column marginals for the grand total (140).

5. In the row labeled x' and the column labeled y', code the class midpoints for the X- and Y-values. In order to simplify subsequent arithmetic, the zero origin should be located near the densest concentration of items. For the present data, then, we would place 0 in the X-interval 50–59 and in the Y-interval 4.0–5.9. These "intervals of origin" may be identified by double ruling of column and row. As the student has previously learned, a plus or minus sign is attached to each coded midpoint according to whether it is larger or smaller than the zero origin.

6. Multiply the coded midpoints by their corresponding marginal frequencies (column f times column y'; row f times row x'); then sum these weighted values. For our example, $\Sigma fx' = 1$ and $\Sigma fy' = -64$.

7. Calculate the weighted squared coded values by multiplying the weighted coded values (column fy' and row fx') by the coded values (column y' and row x'). For our example, $\Sigma fx'^2 = 489$ and $\Sigma fy'^2 = 524$.

8. Multiply each x' by each y' *to obtain the coded cross-products:* $x'y'$s. For example, multiplying $x' = -2$ by each y' yields the following values: 4, 2, 0, −2, −4, −6, −8, −10, −12, −14. These items are conventionally entered in the upper left-hand corner of each cell.

9. Multiply each $x'y'$ by its corresponding cell frequency and place product in lower right-hand corner of each cell. Sum all such items, first adding by rows (columns) and thence to the grand total. For our example, $\Sigma fx'y' = -291$.

10. Substitute sums called for by formula and solve.

that obviate the need to compute slope, the explained deviations, the unexplained residuals, and other quantities involved in the formulas given above. Of course, such mechanical computing methods are not interpretive tools; they are merely recipes that can be followed by any intelligent clerk or machine.

The computing programs here given begin with the product-moment formula, which may be rewritten for somewhat simpler calculation:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \tag{10.3.6}$$

And it may be still more conveniently rewritten, particularly when a desk calculator is available, in terms of the original, untransformed values:

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \tag{10.3.7}$$

This is the computing formula for r, ungrouped data. It intends to eliminate many of the vexatious arithmetic details that are sure to arise when the operation is carried out on the raw deviations around means.

This formula, be it noted, requires only five sums and correspondingly five work columns. Two of these columns comprise the original tabulation of X and Y, so that only three more are needed in order to proceed with the calculation of r. Substituting sums in the formula, we again obtain $r = .89$.

Table 10.3.8 *Worktable for Calculation of r from Ungrouped Values*

X	Y	X^2	Y^2	XY
3	3	9	9	9
7	5	49	25	35
11	7	121	49	77
14	6	196	36	84
15	9	225	81	135
$\Sigma =$ 50	30	600	200	340

$$\begin{aligned} r &= \frac{5(340) - (50)(30)}{\sqrt{[5(600) - (50)^2][5(200) - (30)^2]}} \\ &= \frac{1700 - 1500}{\sqrt{(500)(100)}} \\ &= \frac{200}{224} \\ &= .89 \end{aligned}$$

f		0	4	14	12	25	22	30	23	8	2	140		−64	524	27	318
x'		−5	−4	−3	−2	−1	0	1	2	3	4					−291	
fx'		0	−16	−42	−24	−25	0	30	46	24	8	1					
fx'^2		0	64	126	48	25	0	30	92	72	32	489					
$fx'y'$	+	0	0	3	8	14	0	2	0	0	0	27	−291				
	−	0	23	72	30	19	0	38	70	45	16	318					

$$r = \frac{N\Sigma fx'y' - \Sigma fx'\Sigma fy'}{\sqrt{[N\Sigma fx'^2 - (\Sigma fx')^2][N\Sigma fy'^2 - (\Sigma fy')^2]}}$$

$$= \frac{(140)(-291) - (1)(-64)}{\sqrt{[(140)(489) - (1)^2][(140)(524) - (-64)^2]}}$$

$$= -.59$$

Table 10.3.9 *Correlation Table of Delinquency Rates and Average Monthly Rentals*

DELINQUENCY RATE (Y)	MONTHLY RENTAL (X)															
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	f	y'	fy'	fy'^2	$fx'y'$ +	$fx'y'$ −
18.0-19.9	−35	−28	−21	−14 1 −14	−7 1 −7	0	7	14	21	28	2	7	14	98	0	21
16.0-17.9	−30	−24	−18	−12	−6 1 −6	0	6	12	18	24	1	6	6	36	0	6
14.0-15.9	−25	−20	−15	−10 1 −10	−5	0 1 0	5	10	15	20	2	5	10	50	0	10
12.0-13.9	−20	−16	−12 4 −48	−8	−4	0	4	8	12	16	4	4	16	64	0	48
10.0-11.9	−15	−12 1 −12	−9 1 −9	−6	−3 1 −3	0	3	6	9	12	3	3	9	27	0	24
8.0-9.9	−10	−8 1 −8	−6 2 −12	−4 1 −4	−2 1 −2	0	2 1 2	4	6	8	6	2	12	24	2	26
6.0-7.9	−5	−4 2 −8	−3 1 −3	−2 1 −2	−1 1 −1	0	1	2	3	4	5	1	5	5	0	14
4.0-5.9	0	0	0 5 0	0 4 0	0 6 0	0 3 0	0 4 0	0 1 0	0	0	23	0	0	0	0	0
2.0-3.9	5	4	3 1 3	2 4 8	1 14 14	0 11 0	−1 12 −12	−2 9 −18	−3 1 −3	−4	52	−1	−52	52	25	33
0.0-1.9	10	8	6	4	2	0 7 0	−2 13 −26	−4 13 −52	−6 7 −42	−8 2 −16	42	−2	−84	168	0	136

Unexplained Variation
Regression Line
Line of Least Squares
Product-Moment Correlation Coefficient
Slope
Total Deviation
Explained Deviation
Unexplained Deviation

2. If $r = .3$, how much of the variation in Y is linearly associated with variation in X?

3. Evaluate the following r's as high or low: $-.75$, $.69$, $.25$, $.96$, $-.96$, $.50$.

4. How may one explain the unexplained residuals: as the result of chance or determining factors?

5. Let us suppose that $r = .3$ between college grades and hours of study. Analyze this "low correlation." Is this evidence that study does not affect grades? Discuss.

6. What kinds of marginal distributions should make one cautious about using the product-moment r?

7. Should r be used if the marginal distributions are highly skewed?

8. "The use of r requires that the scatter of cases around the regression line be homoscedastic." Elaborate on this statement.

9. Compute r and r^2 between total expenditures and clothing expenditures (Table 10.2.4).

10. Compute r and r^2 between per capita income and percentage born outside the state (Table 10.2.7).

11. List states in which the percentage born outside is greater than 25. Calculate r and r^2 for this subgroup. Compare this r with that obtained in Question 10 and explain the difference.

12. Compute $b_{xy} = \Sigma xy/\Sigma y^2$ for the illustrative problem given in the text (p. 302). Compare with b_{yx}.

13. What would be the effect on r^2 of greatly extending the ranges of X and Y? (*Hint:* analyze ratio of explained to total variation.)

SECTION FOUR

The Correlation Ratio

Curvilinear Regression. Any index of correlation is a measure of the extent to which one variable may be predicted, or explained, in terms of the other. Of the many indexes that are employed, Pearsonian r is

Function of the Correlation Table. The correlation table should be taken for what it is: a mechanical aid to computation. It is stripped down to essential procedures which help one to get the right answer. It is a kitchen recipe which, when meticulously followed, will guarantee good results even in the hands of amateurs and clerks. Such a recipe, however, does not call for higher criticism of its construction and the selection of ingredients; and the beginning student should not attempt to penetrate these equations and routines for the "meaning" of correlation or for confirmation of its theory — features which are effectively concealed in this "efficiency table."

The device of the correlation table is especially expedient when the research design calls for a large number of correlations or intercorrelations, the computation of which may be delegated to clerical assistants. For this purpose, inexpensive pre-printed commercial forms are available which permit the processing of large-scale operations in orderly manner.

The only precautions which need be observed are those that are relevant to any other tabulation, such as to obtain the optimum size of intervals in order not to introduce unnecessary grouping errors, and to perform suitablo checks on accuracy in calculation.

The execution, by pencil and paper methods, of a correlation chart, even when supplemented by a desk calculator, may seem to some research workers like an uneconomical observance of a traditional device which should be replaced by high-speed computers. There aro three rebuttals to this plausible contention: first, not every worker has a high-speed computer at his beck and call, nor may tho project itself be of sufficient proportions to justify the use of such expensive machines even when available. Typewriters and printing presses have only partially displaced the handwritten record. Second, it is the position of this text that, for the learner, pencil and paper methods, which might seem somewhat primitive to those who cultivate large-scale designs, are still the best laboratory method for the growth of understanding.

Third, such devices as the correlation table are visual aids to the worker; they permit, for example, the examination of scatter to avoid imposing the linear formula upon a too obviously non-linear distribution. These visual aids entirely disappear in the dark recesses of a computer, however efficient they may be to a mature scholar who has the elements of his subject safely behind him, and who may no longer appreciate the aid of the lower rungs of the ladder upon which he has climbed.

Questions and Problems

1. Define the following concepts:
   Coefficient of Determination
   Explained Variation
   Coefficient of Non-determination

Table 10.4.1

Price Index and Marriage Rate, 1887–1906

Year	Price Index (*X*)	Marriage Rate (*Y*)
1887	84	87
1888	87	88
1889	84	91
1890	81	90
1891	82	92
1892	76	91
1893	77	90
1894	69	86
1895	70	89
1896	66	90
1897	67	89
1898	69	88
1899	74	90
1900	80	93
1901	79	96
1902	85	98
1903	85	101
1904	86	99
1905	85	100
1906	88	105

Source: J. M. Reinhardt and G. R. Davies, *Principles and Methods of Sociology*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1932, p. 621 (adapted).

perity, marriage rates will also rise; as prices fall, a symptom of depression, marriage rates will also fall, but not necessarily at a constant rate.

In order to ascertain whether these data conform to the linear or curvilinear model, it will first be necessary to prepare the usual scatter diagram (Figure 10.4.1), which permits us to make such a judgment by inspection. The scatter diagram, on which a freehand curve has been superimposed, suggests the degree and type of relatedness between the two variables. One could have represented the scatter by a straight line, but a curving line follows the contour of the data more closely. We therefore judge η^2 to be a better fit than r^2 and proceed with the calculation of η^2_{yx}.

The procedure for the computation of η^2 is quite analogous to that for Pearsonian r^2. We must calculate the total variation and the explained variation and then find the ratio between them. In short,

$$\eta^2_{yx} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

one of the most conventional, almost commonplace, measures in the field of statistics. But the validity of r rests essentially on two conditions: (1) the relation between the variables must be linear, and (2) the bivariate distribution must be homoscedastic. In general, both of these conditions are provided for when the marginal distributions are normal. Therefore, normality of marginals is often set down as one of the requirements of r.

But these conditions are difficult of fulfillment, for Nature does not necessarily conform to the linear model. Indeed, more often than not the raw data come in non-linear patterns. This is especially true in the social sciences, where many distributions, such as income or size of family, are quite skewed, and the regression lines are therefore markedly nonlinear.

Since r sets up the straight-line model, it becomes inappropriate to the extent that the relation between the variables departs from linearity; and also as the scatter diagram becomes heteroscedastic, even though it maintains linearity. To be sure, since these ideal conditions can never be realized, statistical workers are, by dint of circumstance, necessarily tolerant of approximations. But there is a prudent limit beyond which approximations should not be permitted to extend, especially if more fitting measures are available.

There are various devices with which curvilinear relations may be measured, only one of which will be treated here. The method set forth is the *correlation ratio*,* or η (the Greek lower-case letter *eta*), but better expressed as η^2. It is a relatively simple procedure and adheres to the same principle on which r^2 is founded, namely, the ratio of explained to total variation. It differs from r^2 procedurally in that the explained variation is derived from the column (row) means of the correlation table instead of the hypothetical linear regression line. The correlation ratio, η^2, thereby supplements Pearsonian r^2, which latter enjoys a vogue and repute that are probably greater than its utility in the realm of social data justifies.

Calculation of the Correlation Ratio. For simple illustrative purposes, we shall compute and analyze the correlation ratio between price index and marriage rate, 1887–1906 (Table 10.4.1), data of manageable volume.† Here, the sociological hypothesis is this: as prices rise, indicating pros-

* Conventionally, "correlation ratio" is used synonymously with η. In the course of the discussion, it will become clear that η has little meaning for all practical purposes; it is η^2 that is the meaningful measure of correlation. Hence, the term "correlation ratio" will be used interchangeably with η^2.

† Since these data are in the form of a time series, we would normally test them for lag between the two variables in order to determine whether the data should be paired as given or with a lag of a year or more, on the assumption that marriage rates would not react to prices until after a lag. This analysis will be dispensed with here.

Table 10.4.2 *Worksheet for Calculation of* η^2_{yx}

MARRIAGE RATES (Y)	WHOLESALE PRICES (X)								
	65-69	70-74	75-79	80-84	85-89	f	y'	fy'	fy'^2
105-109					1	1	3	3	9
100-104					2	2	2	4	8
95-99			1		2	3	1	3	3
$\bar{Y}=93$									
90-94	1	1	2	4		8	0	0	0
$\bar{Y}_c$									
85-89	3	1		1	1	6	−1	−6	6
f	4	2	3	5	6	20		4	26
$\bar{Y}_c$	88.25	89.50	93.67	91.00	98.67				
$\bar{Y}_c-\bar{Y}=d$	−4.75	−3.50	.67	−2.00	5.67				
fd	−19.00	−7.00	2.00	−10.00	34.00				
fd^2	90.25	24.50	1.33	20.00	192.78	328.86			

$$\bar{Y} = \bar{Y}' + \left(\frac{\Sigma fy'}{N} \times i\right)$$
$$= 92 + (\tfrac{4}{20} \times 5)$$
$$= 93$$

5. Calculate the sum of squares in Y, or *total variation*, by the computing formula (see p. 166):

$$\Sigma fy^2 = i^2\left[\Sigma fy'^2 - \frac{(\Sigma fy')^2}{N}\right]$$
$$= 5^2\left[26 - \frac{(4)^2}{20}\right]$$
$$= 630$$

6. Compute the mean of the Y-values within each column, $\bar{Y}_c$, and insert the result in the row carrying that designation. Thus, the mean of the first column would be:

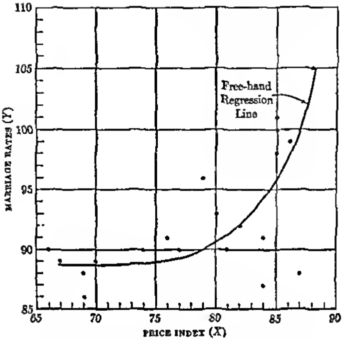


FIGURE 10.4.1 *Scatter Diagram, Wholesale Prices and Marriage Rates, 1887–1906*

However, the details of the procedure vary. This is attributable to the fact that the "line of regression," which supplies us with the explained deviations, is now a line of column means instead of an unbroken straight line. Accordingly, we must determine the mean of the values in each column (row) in the correlation table, the grand mean of these values, and the deviations between the respective column means and the over-all mean. The entire procedure is best described in terms of the following steps:

1. Employ the familiar procedure of grouping bivariate data into intervals of proper width — 5 in this instance — with Y-intervals in rows, X-intervals in columns. The result is shown in Table 10.4.2.

2. Determine the marginal frequencies (f) and check the grand total.

3. Set up code columns: y', fy', fy'^2, as in the previous correlation chart.

4. Calculate the grand mean of the Y-values by Formula 5.3.3, in which Y' = the midpoint of the class interval which has been coded zero, and i = class interval.

Table 10.4.3 *Worksheet for Calculation of η^2_{xy}*

	WHOLESALE PRICES (X)									
MARRIAGE RATES (Y)	65-69	70-74	75-79	$\bar{X} = 78.75$ 80-84	85-89	f	$\bar{X}_r$	d	fd	fd^2
105-109					1	1	87.00	8.25	8.25	68.06
100-104					2	2	87.00	8.25	16.50	136.12
95-99			1		2	3	83.67	4.92	14.76	72.62
90-94	1	1	2	4		8	77.62	−1.13	−9.04	10.22
85-89	3	1		1	1	6	73.67	−5.08	−30.47	154.84
f	4	2	3	5	6	20				441.86
x'	−3	−2	−1	0	1					
fx'	−12	−4	−3	0	6	−13				
fx'^2	36	8	3	0	6	53				

$$\bar{X} = \bar{X}' + \left(\frac{\Sigma fx'}{N} \times i\right) = 82 + \left(\frac{-13}{20} \times 5\right) = 78.75$$

$$\Sigma fx^2 = i^2\left[\Sigma fx'^2 - \frac{(\Sigma fx')^2}{N}\right] = 5^2\left[53 - \frac{(-13)^2}{20}\right] = 25[44.55] = 1,113.75$$

$$\eta^2_{xy} = \frac{\Sigma fd^2}{\Sigma fx^2} = \frac{441.86}{1.113.75} = .40$$

variation within rows and columns will not be exactly equal; hence, η^2_{yx} will generally not be equal to η^2_{xy}. In the extreme — but very unlikely — case, there could be absolutely no variation within columns but a substantial amount within rows, so that η^2_{yx} would be unity but η^2_{xy} only moderately high.

Conditions Under Which η^2 May Be Employed. The conditions under which η^2 is applicable are more liberal than are those of Pearsonian r^2. No restrictions are placed on the pattern of marginal distributions, which

Midpoint (Y)	f	fY
92	1	92
87	3	261
	4	353

$$\bar{Y}_c = \frac{353}{4} = 88.25$$

7. Find the deviation of each column mean from the grand mean: $\bar{Y}_c - \bar{Y}$. *In the first column:*

$$88.25 - 93.00 = -4.75 \text{ (Row "}d\text{")}$$

8. Weight each deviation (d) by its corresponding column frequency and *insert in proper row* (fd). *In the first column:*

$$-4.75 \times 4 = -19.00$$

9. Square each *explained deviation.* Operationally, this merely *involves multiplying d by fd, or adjacent rows. Again in the first column:*

$$-19.00 \times -4.75 = 90.25$$

10. Find the sum of the weighted squared deviations:

$$\Sigma fd^2 = 328.86$$

This is explained variation; it is the variation in Y that would have been observed if each Y-value were located exactly at the mean of its own column — as if there were no variation within columns.

11. Divide the explained variation by the total variation, which is the correlation ratio as defined above:

$$\begin{aligned}\eta^2_{yx} &= \frac{\text{Explained Variation}}{\text{Total Variation}} \\ &= \frac{\Sigma fd^2}{\Sigma fy^2} \\ &= \frac{328.86}{630.00} \\ &= .52\end{aligned} \qquad (10.4.1)$$

Thus, we find that 52 per cent of the variation in marriage rates is explained by variation in price indexes.

In exactly analogous fashion, but operating on row means instead of column means, we may determine η^2_{xy} — the proportion of variation in X explained by Y. Upon completing this operation, which is compactly summarized in Table 10.4.3, we find that $\eta^2_{xy} = .40$.

The fact that these two η^2-values are different illustrates the significant principle that, unlike r^2, η^2 is not reversible. In general, the relative

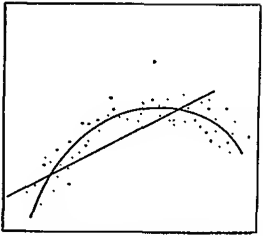


FIGURE 10.4.2 *Curvilinear Scatter*

the curve, since the straight line becomes less and less representative as the data become more and more curvilinear. In fact, on the edges of the data, the straight line departs increasingly from the swarm as one projects the prediction beyond the observed values.

Furthermore, the norming potential of r^2 is reduced when applying the linear model to curvilinear reality because the index cannot range from zero to unity. It will be recalled that ϕ_{max} was less than unity whenever the marginal distributions were not identical. Analogously in this case, an r^2 of unity would be unattainable when mistakenly applied to non-linear data. But η^2, as already shown, is not so restricted, and could theoretically attain unity. In the case of ϕ, it was suggested that under certain circumstances we might resort to Q; similarly, when r^2 is inadequate, *we turn to* η^2.

Precaution in the Application of η^2. In order to implement the formula of η^2, it is necessary to group the data into segments, and such grouping always involves a certain degree of arbitrariness. By pushing this process to an absurd extreme, we could conceive of as many class groupings as there are points, with each observed point representing its own column. The regression curve would then pass through every point. Under such an arrangement, η^2 would be perfect unity since no deviation within the columns would be possible. On the other hand, groupings could be made absurdly large, and thereby fail to etch the pattern of distribution. In such a case, the residuals would be too large to do justice to the curvature of the pattern, and once again the familiar phenomenon of large grouping errors would appear.

The general rule is to construct the columns of such width that the means of the columns are fairly stable, which implies that the means may be normal, skewed, or even bimodal; the regression line may follow any curvilinear pattern. Even a qualitative variable may be used on one axis, although not on both (If both were qualitative, there could be no arithmetic means, and we would have to return to the contingency coefficient, C.) Only one limiting factor may be suggested: the plotted observations should, in general, be homoscedastic at least in one direction; and this for the same reason that Pearsonian r sets up the same requirement, namely, a synoptic measure is usually more meaningful when made up of homogeneous items. *But this is a logical reason, not* a mathematical restriction at all It is the same reason that sometimes renders a mean of widely dispersed data inadvisable.

Principles of Interpretation. Since the hypothetical regression line shuts its direction as a result of its curvatures and bends, one cannot interpret η as one would r: "the higher the X, the higher the Y" (technically, the average change in Y per unit of X). Such an interpretation would be meaningless, because the regression line may change direction; henre, *η is not a counterpart to the constant slope, r. Indeed, η has no definable* function at all in the measurement of correlation; it is merely the square root of η^2, which is the proper measure of curvilinear correlation.

Since each swarm of plotted points will always have a certaln amount of curvilinearity in it, η^2, which reflects curvilinearity and linearity equally well, will therefore always be a more precise and complete statement of explained variation, and therefore will always be higher than r^2. It will always maximize our statistical explanation because it adapts itself automatically to the configuration of the data. On the other hand, r^2 is more rigid, *restricting* its potency to reflecting the linear *model. When* perfect linearity prevails, $\eta^2 = r^2$.

The reason η^2 is able to reflect the explained variation *more* fully is simply that it in effect divides the curved regression line into segments which will always be in closer proximity to the points In the swarm than would be a single straight line. Hence, the residuals will be smaller. In a strictly linear swarm, with only one direction, this flexibility is, of course, not called for.

In Figure 10.4.2, it is obvious that the straight line does not do justice to tho swarm as does the curved regression line. However, there is nothing in the matbematical processes which will make it impossible to use the lincar formula on curvilinear data. It will "pick up" whatever linearity there is — which may not be very much. A curved line is simply a better fit and therefore more adequate for our purposes.

This better fit becomes all the more evident when we use the regression line for its intended purpose, namely, the prediction of one variable from the other. It is quite apparent that a Y-value read from the straight line will be subject to much larger error than it would be when read from it. Specifically, our questions are: (1) what is the essential meaning of correlation? (2) how competent are the statistical tools to measure correlation? and (3) how can one best interpret a correlational measure after it has been obtained?

The Meaning of Correlation. It would be naive to take for granted that any obtained index of association between two variables is necessarily a measure of the true relationship between them. Nature does not reveal herself so readily in response to the application of a mere man-made contrivance such as a formula or equation. These natural relationships are, after all, quite complex: they may be strong or weak, intimate or remote, simple or complex, genuine or spurious.

When statistical linkages are analyzed, we find that they may be any one of several general types. These will be briefly reviewed.

(1) *Chance Relation.* By definition, a "chance relation" is, of course, a contradiction in terms. There may be a chance joint occurrence, but no "chance association." For, if there is any relation at all, the linkage or association is represented by a frequency which exceeds chance. Nevertheless, the expression "chance relation" is often informally and conveniently used to identify the frequencies which exemplify statistical independence between variables. *Strictly speaking, therefore, chance relation is the negation of relation, statistically expressed as zero correlation;* it thereby sets the lower boundary to our descriptive vocabulary for the complete range of relational measures.

(2) *Nonsense Correlation.* An obtained measure of relation may be said to be a sheer coincidence when it does not represent a "true" functional relation. We may select any set of observations, correlate it with any other set, and fortuitously obtain a "correlation."

To illustrate: it may be shown that as the per capita consumption of liquor has increased in the United States the average length of life has also increased. Does this indicate a positive relation between the intake of liquor and length of life? Does this contradict the stereotyped opinion that liquor shortens life? Statistically, the correlation is impeccable; but abstract statistics are indifferent to social content. When the social content is analyzed, not only do the two series of data seem quite remote, but they do not even necessarily refer to the same members in the population. The individuals who "drink" may not be the individuals who live long. It is quite true, people in general are living longer than ever and liquor is also being consumed in ever greater quantities. But these two sets of data may be said only to coexist, not to interact. Yule dubbed such mere arithmetic "correlations" *nonsense correlations.* They are frequently cited by those who wish to disparage statistical methods by showing what absurd propositions can be "proved" by statistics.

(3) *Correlation as Evidence of Cause and Effect.* In layman's language,

would not vary appreciably were the boundaries of the intervals narrowed or extended. *Also, if the plotted points* thin out on the edges of the data, it is good policy to eliminate these border data since they tend to increase disproportionately the over-all correlation.

Questions and Problems

1. Does the calculation of η^2 assume a smooth regression line?
2. In what respect does the correlation chart differ from the simple frequency table?
3. Would it be possible to calculate η^2 between two qualitative variables? Why or why not?
4. Is the numerical value of η^2 affected by the order of columns (rows)?
5. Under what tabular conditions would r^2 and η^2 be equal in numerical value?
6. Show graphically the conditions under which r would be approximately zero and η^2 1.00.
7. *Verify graphically that* η^2 *may be brought close to unity by employing as* many columns as there are values.
8. When the regression is curvilinear, why is homoscedasticity not likely to be *uniform in both directions? Show graphically.*
9. Calculate η^2_{yx} and η^2_{xy} between delinquency rates and average monthly rentals (Table 10.2.2).
10. Calculate r^2 between price index and marriage rate and compare with the obtained η^2-value of .52 (Table 10.4.1).

Section Five

Some General Guides to the Interpretation of Correlation

The skillful computation of an abstract index of correlation does not exhaust the responsibility of the sociologist, nor is the calculated result — which is often more or less routinely obtained — a guarantee that its sociological significance has been grasped. It is quite possible to be thoroughly familiar with the purely statistical character of a correlational index and still be completely unaware of its sociological import.

All statistical correlation rests on the assumption of some kind of *cohesion in nature, whether observed directly or under experimentally* controlled conditions. We are very much concerned, therefore, about just what it is that correlation is supposed to measure, and how well it measures mistaken, the inference may be revised. Such a procedure does not differ in principle from that practiced by an auto mechanic in locating the "cause" of the rattle in an old car, or the current procedure in determining the cause, or causes, of colds, cancer, or of war.

(4) *Correlation as a Measure of Common Factors.* The high correlation which we could expect between economics and sociology grades could not be primarily explained by a causal relation between the two, but rather by the common factors which make for academic achievement in these two related subjects. The grades achieved in economics and sociology are merely two slightly different manifestations of almost exactly the same phenomenon, the components of which are: degree of intelligence, habits of study, grading systems, similarity of subject matter, common motivation and interest. It would be tautological to state that the two sets of data, being made up of the same content, would vary together. To the extent, however, that other factors do intrude (for example, variation in teaching staffs, subject matter, and motivation), the correlation would be correspondingly reduced. High delinquency and high truancy may similarly be associated, not because one causes the other, but because they too, like sociology and economics marks, are characterized by common factors: poverty, unsupervised play, absence of mother from home, and the like. In the field of heredity, the characteristics of siblings, which show many similarities, are likewise the resultant of common antecedent hereditary factors. In fact, this was the original conception of correlation as propounded by Francis Galton, who first gave currency to the concept through his studies in heredity during the last quarter of the nineteenth century.

The various ways in which two measured variables might be bound together by underlying common factors may be schematically represented by means of two intersecting circles. The presence of hypothetical common factors is indicated by the overlapping areas of the figures. The first diagram would represent the case of correlation between economics and sociology grades which are compounded to a considerable extent of identical ingredients. The second illustration pictures the very high correlation coefficient between test-retest scores on a social attitude inventory, administered to the same population before and after an intervening period. Diagram 3, in which there are no common elements, portrays a near zero correlation, an example of which might be the relation of hair color and susceptibility to disease. Actually, it is quite difficult to find a convincing illustration of genuine zero correlation; for, if underlying factors are pursued extensively enough, some remote interconnecting common factor will doubtless be dredged up. The last figure, in which one variable is totally included within the other, is again an instance of a hypothetical possibility, but probably never in complete correspondence with fact. A generally humane attitude on public affairs would dictate a person's as well as that of professionals, there is no concept so commonly employed, and so useful in the understanding of the world about us, as the concept of "causation." We speak of the causes of death, the causes of accidents, the causes of divorce, of delinquency and crime. It is thought that, if the causes of delinquency, disease, and war could only be uncovered, we could control their appearance. Statistical researches in biology are built around correlations that are designed to reveal, for example, the effect of nutrition on plant and animal growth.

The causal influence may flow unilaterally in one direction; or bilaterally in both directions, simultaneously. Thus, the hours of study may "cause" good grades, and good grades may also encourage more study; higher wages may cause higher prices, and higher prices bring higher wages in their famous spiral effect. Childlessness may influence divorce, but the prospects of divorce may cause childlessness; marriage may increase length of life, and healthy, long-lived people tend to marry.

In spite of the wide prevalence of the concept of "causation" and its quite obvious utility, there is in some intellectual circles a fashionable objection to it. The notion that one variable may exercise a force upon another seems to some thinkers too mystical to be credible. According to one school of thought, all we can know is empirical association and sequence. As the argument of Hume runs, we cannot perceive causation, but only statistical contingency; the only statement we may make is one of measurement of invariant relations between massed observations.

According to this view, therefore, a statistical measure of correlation *does not necessarily indicate causation.* We may state only that correlations reveal a mathematical relation, in the sense that values vary with one another. Consequently, a rugged empiricist will leave the question open as to whether there is such a thing as causation in the first place, and if so, whether its direction can be ascertained by statistical means.

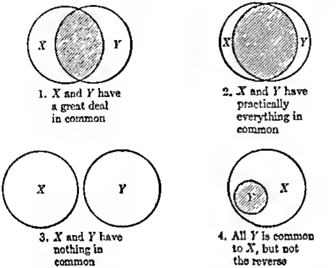
Although these sophisticated questions are of undoubted fascination, common sense will not renounce the concept of cause. No amount of intellectual equivocation will raise doubts in the mind of a gardener that *for all practical purposes* rain "causes" the plants to grow and not the other way around. An intellectual who may reject the concept of cause will nevertheless in real life act as though it existed. The applied scientist, who is called on to manipulate and control, will use the conception of causative factors as an indispensable guide to action. Statistical correlations are undoubtedly an aid in discovering, if only by inference, where causation operates, and how strong it is.

The practicing statistician is therefore not primarily concerned with the metaphysics of causation, even though the subject may delight him in his leisurely intellectual moments. If, after various statistical computations, one or more variables seem to produce another, *we label those factors as "cause."* If, after further observation and experimentation, we are provided for in the formula. Things would be simple if that were so. The difficulty is that, in addition, this measure will reflect an unknown number of variables which are not represented in the formula. This is due to the well-known circumstance that there are a large number of factors involved in each event that are constantly interacting and operating through one another but for which there is no room in the formula. We must realize that a formula may accommodate relatively few variables, according to its design and make-up. But by restricting ourselves to the observation of only a few variables, we do not immobilize or exclude from operation, within the field before us, the numerous other factors which continue to produce their effects on our measures whether or not we happen to be looking at them. For example, when we measure the height or weight of a person, we are also measuring the effects of his age, sex, race, lifelong nutrition, way of life, and even the time of day (since height normally shrinks during the day, and weight increases). These factors which intrude in our measurements are called *concealed factors*, since the surface information carries the effect of such hidden factors without identifying them. In a correlation between race and death rates in the United States, the measure would indicate a high correlation, since Negroes have a high death rate and whites a low rate. However, in measuring Negro death rates, we are also unwittingly measuring the death rates of the lower socio-economic levels to which the Negroes predominantly belong. A classification of the United States population by race is, to a large extent, at the same time a classification by socio-economic background. By corollary, we are measuring the death rates of higher socio-economic levels when nominally measuring rates of the white race. Hence, we cannot take the correlation at its face value. The variable race is spuriously credited with the cause of death rates, whereas the concealed factor of socio-economic status is the "true" factor with which death rates are associated.

To be sure, every correlation contains some effects of concealed factors that are doing the work for which the *spurious factor*, named in the correlation, is superficially credited. A *spurious correlation* may be defined, therefore, as a correlation in which hidden factors are exerting the influence for which the surface factor is erroneously credited. Clearly, it is the interpretation that is spurious, rather than the statistical correlation as such.

Spurious correlation is often employed as a derogatory epithet — as though a correlation, computed from a small fragment of our complicated web of life, could ever be completely cleansed of concealed factors. Sometimes, however, the degree of spuriousness is absurdly obvious, bordering on nonsense. Karl Pearson first proposed the concept *spurious* in 1897, and Yule, as previously noted, termed correlations *nonsense correlations* when the remoteness or absurdity of the association between the identified variables was apparent to any thinking person.

Since there is a degree of spuriousness in every obtained correlation,



1. *X* and *Y* have a great deal in common

2. *X* and *Y* have practically everything in common

3. *X* and *Y* have nothing in common

4. All *Y* is common to *X*, but not the reverse

Figure 10.5.1 *Correlation as a Measure of Common Factors*

attitude on a specific issue, in which case the part would be included within the whole, without remainder. However, human attitudes are not quite as consistent as all that, for everyone shows some symptoms, however slight, of a schizoid character.

A type of correlation which at first glance may appear to be embodied in the foregoing classification is sometimes labeled "part-whole," but should more realistically be described as "self-correlation." It diverges from the preceding system in that the "common" elements derive from the fact that the observations are, in part, duplicated in the two series of data. In a comparison of freshmen and all-university grades, the freshmen grades appear as duplicates within the total university observations. To the extent that freshmen are a smaller or larger component of the respective student bodies, the resulting correlation will automatically be lower or higher, for the very evident reason that correlation between identities is bound to be perfect. The result will be a truism rather than a revelation. It is even conceivable that the "part" will show little correlation with the remaining components and yet, because of the size of the common part, yield a considerable, but redundant, correlation. It raises a question whether such a computation should be undertaken at all, and whether it would not be more prudent to correlate the individual components among one another.

(5) *Spurious Correlation.* An index of correlation presumably measures the degree of relationship between the two (or more) variables which are

constitute nature and society, and it ignores the useful services which these concepts have already administered. Far from being supernumerary, the variety of correlational measures might with better logic be construed as evidences of adaptability to the intricacies of nature.

Basically, the suitability of a specific correlational formula is judged according to the following three general criteria: (1) the pattern of relationship between the specified variables; (2) the nature of the data, or the form in which they are expressed; and (3) the desired degree of accuracy which is differentially yielded by the equally available measures. Each of these criteria will be discussed in its essentials only, for it is not necessary at this stage to reiterate the detail now familiar to the student.

(1) *The Pattern of Relationship.* The multiplicity of indexes derives, to a very large extent, from the fact that "relationship" is not a monistic concept; hence, no single index could ever presume to be omnicompetent in the presence of the various types of configuration of the interrelationship. Thus a linear relation is distinguishable from curvilinear interdependence: a linear relation is defined by a uniform rate of change in one variable for every unit of change in the other; while a differentially changing rate characterizes the curvilinear relation. A relation between only two variables is not identical with that between three or more variables, since, as the number of variables increases, the network which connects them becomes more intricate. Relations may also have a directional dimension, conceived as one- and two-way dependence. A detailed familiarity with the respective formulas and their derivation will reveal the subtler differences between the various connotations of "relationship" which the symbols are designed to express.

(2) *The Forms of the Data.* Data do not always come in identical forms and consequently are not measurable in the same units. Attributes are enumerated in frequencies, and variates are expressed in magnitudes. Since a formula must have the capacity to digest the data which are fed into it, we will require in general as many types of formulas as there are different forms of data.

For example, attributes are presented in $r \times c$ contingency tables, and their association is measurable by the degree of discrepancy between chance and observed joint frequencies. Quantitative variables, on the other hand, are presented as magnitudes whose correlation is measured by the concomitant variation symbolized, for example, as r and η.

(3) *Degree of Precision Required.* At his present stage of development, the student's statistical mentality should recognize that nature cannot be clearly duplicated on paper in the shape of formulas and symbols. Therefore, any formula is but an ideal-typical construct which only approximates, more or less, the raw, fragmentary data torn out of context. Thus, Pearsonian r assumes that the bivariate data are distributed normally; ϕ works best when the marginal sets are identical. But nature in the raw never

the concept loses its characteristic as an epithet. It is a foregone conclusion that social behavior is complex, and that no correlation is actually what it seems. Depending upon one's detailed interest, every correlation should be analyzed and broken up, and could be pursued indefinitely to greater refinement. No one, for example, would reject or undervalue the association between moisture and plant growth because of the many *intervening variables* which actually "do the work" being measured in the correlation: between moisture and its ultimate effect are such factors as the solution of chemicals in the soil, temperature, proximity of other plants which compete for the moisture, and sunlight.

Is there, then, any routine test by which we can tell whether a correlation is spurious or genuine? Actually, as has already been implied, no correlation is wholly genuine in the sense that hidden factors are completely silent. But statistical devices are available which will "eliminate" one or more of the spurious factors and lay bare the "genuine" factors, thus measuring their separate effects. Simply put, this involves introducing additional variables into the system, and testing them for correlation. In general, if the additional variable incorporated into the computation changes the original index, or leaves it unaffected, we may state, respectively, that the original correlation was spurious or genuine.

More specifically, there are three relatively simple, interrelated methods by which additional variables may be incorporated into the measure of relation, two of which have been previously discussed. (1) The method of norming by subclassification has been presented as a tool for testing the genuineness of an observed relation (Chapter 7, Section 2). (2) Standardization serves additionally to summarize the specific classes of observations into a single measure, by which we may test the validity of the relation between variables. (3) *Partial correlations* are designed to measure the correlation between two specified variables while one or more others are statistically held constant. The technique of partial correlation, which is not presented in this text, is merely an extension of the Pearsonian r. In general, partial correlations require more elaborate computations, but invoke no new principles or concepts.

Why So Many Types of Correlation Measures Are Required. When confronted with the long list of correlation measures — Q, ϕ, ρ, C, r, η, as well as numerous others not represented in these chapters — a student might well become skeptical of the validity of a concept which yields so many different answers to the same question: "How strong is the affinity between two (or more) sets of data?" Do they all possess the same degree of competence to answer this question? Does not the multiplicity of answers to the same question tend to discredit the validity of them all?

Such healthy though immature statistical agnosticism does not evince adequate appreciation of the extremely intricate web of relations which there are certain typical pitfalls into which students may stumble, and some of these are here itemized as aids in interpretation. These hints may be analyzed and illustrated by the student, and will go far toward enriching his understanding of correlation in general, and of social science as well.

(1) *Time Lag.* Two series of data, such as prices and wages, may be very definitely linked, each influencing the other. However, the influence of one on the other may occur only after a lag of a period of time. Consequently, to measure the inherent relation between the two time series, it would be incorrect to pair the data within the same year; rather, we should pair them diagonally, at intervals of a year or two, depending on the estimated lag. After a little experimental pairing, the most plausible time lag will be revealed by the highest attainable correlation. Such a syncopated pairing would not, of course, be justified when the time factor is not functionally involved.

(2) *Measures Applied to Heterogeneous Data.* The non-uniform scatter (e.g., gourd-shaped) around the line of regression is evidence of heterogeneity of the data — a distribution that is called *heteroscedastic.* Since a wide scatter indicates low correlation, and a narrow scatter high correlation, the resulting *r* does not do justice to either extreme. Such an *r* would be analogous to combining in one mean the salaries of janitors and managers, with a consequent equal representation of both groups. Under most circumstances it would be far better to segment the data into more homogeneous groupings to obtain a measurement which would do justice to the high (or low) correlation in the respective subgroups. This is not an illegitimate procedure for the purpose of pushing up the correlation measure. Quite to the contrary, it is an intelligent procedure in the analysis of data to arrive at more realistic and serviceable indexes.

(3) *Specificity of an Index.* We must never extend the generalization beyond the specific context in which the data are found. It would be an error to overgeneralize in the following manner: "The correlation between delinquency and monthly rentals is −.59." It would be more accurate to say: "The correlation between rentals and delinquency, in a particular city, at a given period, and under given conditions was −.59." If, for example, new definitions of delinquency were to be imposed, or if rent-control laws were modified, or if any one of a dozen relevant factors in the social setting were modified, the correlation would have been altered. It is this circumstance that should warn us against quoting or applying obsolete correlations as though they pertained to a current setting. Correlations, like the folkways of society, are culture-bound. Extrapolations are therefore always hazardous.

This caution will probably elicit considerable methodological discussion on the cumulative character of social research, as compared to the cumulative potentialities of physical laboratory science, where the contextual satisfies such ideal assumptions. This being the case, why not invent procedures which are grounded in "reality" rather than in ideal conditions which admittedly do not exist? The answer is that it is impossible to invent a formula for every conceivable set of contingencies or for every possible gradation in their forms. Just as the law of falling bodies is standardized for a vacuum, so the correlation formulas are standardized for ideal conditions which, in fact, do not exist. What is a still greater violation of nature is the necessity of segmenting and segregating the objects of study, and abstracting them from the raw context in which they normally operate.

Considered from this point of view, a formula is a human contrivance like a machine, which is supposed to work for us. We may ask it to work on materials for which it is not intended. This would be, of course, a blunder on our part, ascribable to deficient experience and knowledge. Or, what is more likely, we may put in materials for which it is only moderately well suited; or again, we may assign it a task for which it is eminently well suited.

What we are saying is that precision is relative to our needs, our interests and the potentialities of our measuring tools, and that occasionally it may be a matter of selecting one of several alternative formulas. Thus, r and η are to a slight degree interchangeable, depending on the degree of tolerance on the part of the worker; ρ is similarly acceptable for r; ϕ is another form of r with possible alternation between them.

Convertibility of Indexes. At times, we may have left the impression that the various formulas were mutually exclusive; however, it has also been intimated that some of them possess certain elements in common. Thus, the ϕ-coefficient may be viewed as the r formula tailored to dichotomous data; η^2 and r^2 are identical measures when the relation is perfectly linear; ρ, based on ranks, is also a derivative of Pearsonian r. Such reflections testify once again to the ingenuity exercised in the invention as well as in the implementation of formulas in our effort to penetrate the secrets of nature.

Pitfalls in the Interpretation of Correlations

It has been repeatedly emphasized that a correlational measure should not necessarily be taken at its face value. For example, a Pearsonian r of zero does not prove the absence of a substantive relation between the two series of data: we may have employed the wrong formula; an r of .3 might be substantially raised or lowered by subclassification. How can we ever be sure of the validity of the index? The answer is, of course, that we may never be absolutely sure. It is necessary to be pragmatic and interpret the results in terms of relevant knowledge. At the same time, under study, although his memory and understanding may accommodate a very large number of variables, depending upon his experience, intelligence, and opportunities for observation. It is therefore highly probable that on many occasions an experienced worker will, in a certain sense, know much more than his formula does. He can attend to a larger number of variables, but probably not quite so accurately as can his mathematical tools. A good worker will always combine and blend the potentialities of his mathematical tools with the subtlety of his own imagination.

What Is a "High" Correlation? A correlation is judged "high" or "low" according to the same standards as any other phenomenon is so judged: either in absolute terms, or relative terms. In absolute terms, a correlation is high or low as it approaches its possible numerical limits. Hence, .9 is high, and .2 is low. However, in terms of the norms of human expectation and demand, absolute terms are not very meaningful. A correlation of .7 would be considered very low for test–retest problems, where the reliability of the test is in question, but very high for the correlation between *IQ* and grades. Familiarity with the scientific norms prevailing in given situations offers the only guide for useful judgments.

Are Low Correlations Acceptable? Is a zero correlation a satisfactory result in sociological inquiry? One may assert that science is impartial and is not interested in any particular findings, and hence a zero correlation is as satisfactory in principle as a high correlation. It is true, the scientific attitude is not biased in favor of any particular finding. However, there is a difference between a "finding," and a scientifically acceptable correlation measure which answers in part the question which prompted the investigation in the first place. Science has as its objective the completest possible explanation; a zero correlation explains nothing.

This is not to say that zero correlations may not have their uses, for they may explode long-cherished myths and eliminate tentative working hypotheses. But constructively, they represent failure, for they add nothing to human knowledge. Humanly speaking, zero correlations represent statistical independence; they permit no prediction beyond chance, which is the ultimate of ignorance, rendering adaptation or survival difficult or impossible. In a random world, we could never learn by experience.

It is one thing to say that science is not biased in favor of any specific finding (that is, any specific prediction) but another to state that science is not interested in prediction at all (that is, accepts zero correlation).

QUESTIONS AND PROBLEMS

1. Distinguish between nonsense and spurious correlation.
2. Discuss in what manner knowledge of subject matter will influence the interpretation of obtained correlation measures.

circumstances are more constant. The potential non-comparability of social data is admittedly a persistent barrier to progress in the social sciences. The United States Census, one of the most important sources of quantitative data on a large scale, has modified the definitions of family, employment, occupations, urban areas, and other basic concepts at intervals of ten years, which is one formidable example of the relevance of this principle of specificity.

(4) *Correlation Between Synoptic Measures.* The data for correlations are frequently in the form of synoptic measures, such as means and percentages. The mean, for example, has the familiar effect of automatically eliminating all variation, and thereby reducing error of an undetermined amount. Since variation is a form of error, its reduction or elimination will always tend to raise deceptively the coefficient of correlation over what it would have been had the original itemized data been used. Special care must therefore be exercised in interpreting correlations of synoptic data.

Thus, percentages may be a source of misinterpretation. The percentage of illiterates by states, correlated with the percentage of Negroes, would yield a rather high index. It is, however, conceivable that a large proportion of illiterates in each state are rural white, and this would tend to contradict the impression of the correlation between race and illiteracy. A more forthright procedure would be to cross-tabulate individuals by race and literacy. If such detailed data are not available, it still devolves upon the investigator to exercise imaginative restraint in his interpretation.

(5) *General Sociological Considerations.* All general precautions observed in ordinary statistical procedures must also be taken in correlational computations. Uniform definitions of terms, units of measure, and care in collection of the data are prescribed for statistical workers of every type. But the specific pitfalls vary for different types of computations. Correlations, especially when used for comparative purposes, may be entirely misleading when definitions of items are vague and unreliable. Thus, a compilation of crime rates in different areas of the United States, where definitions of crime as well as court procedures vary, might be fairly meaningless unless such definitions were standardized. A correlation between monthly rentals and delinquency rates in places where rent control prevails would yield a different measure from what we would find in an area or chronological period where rents are determined by normal competitive economic processes. It is clear that no correlation is "terminal" but is susceptible to continuing analysis, refinement, and extenuating interpretation. To carry out such functions requires not only a competent statistician in the mathematical sense, but also a worker competently trained in the intricacies of the social forces operating in the area of observations.

An investigator must never lose sight of the obvious fact that a correlation formula usually reflects only a few of the variables in the context

Sampling 11

The Sampling Attitude

The Purpose of Sampling. The broad aim of statisties is to describe and summarize mass phenomena like births, deaths, and income, and their interrelationships. However, it is often necessary or practicable to hase such description on a fraction of the total aggregate, and sometimes an exceedingly small one at that. Such an expedient may be, and usually is, quite satisfactory. It is astonishing how effective a well-selected fmgment can be: a small snippet from a bolt of cloth; a few drops of blood from the patient's total supply; a few thousand survey votes, hy which we describe the political intentions of millions of voters. Such procedure is standard practice in everyday social and economic life, as well as in the hranches of scientific activity. In instances of this kind, when the data *are partial rather than complete, and when they are used to characterize* the entire set, we call the fmgment a *sample,* and the total aggregate n *universe,* or *population.* We name a specified value of the universe, such as the mean, a *parameter,* and its counterpart in the sample we term a *statistic.* The objective of sampling is, therefore, to dmw an inference ahout the parameter, which is unknown, from the sample statistic which is observed. This process of genemlizing in a prescribed manner from sample to universe has come to be known as *statistical inference.*

Although statistical inference as a formal quantitative technique is principally an achievement of the twentieth century, its underlying motivation is as ancient as mankind itself. In his continuing effort to adjust to the environment, man has necessarily been obliged to treat an isolated experience as typical of a larger system in order to profit hy it. Once stung by a bee, or hurned on a stove, he is likely to view such objects as persisting hazards to be avoided. On the same principle, hut in a more deliberate manner, he samples one peach from the bushel, plugs a watermelon, or gives the new car a trial run.

3. Illustrate, from your own experiences, if possible, the concept of nonsense correlation.

4. Give several illustrations of spurious correlation, and show how the spurious element may be isolated.

5. Itemize the different conceptions of correlation, and show how they are interrelated.

6. Define and illustrate a concealed factor, and show how it may deceive an observer.

7. Explain the general principles which justify the use of various measures of correlation.

8. Define a synoptic measure, and show how it may create problems in the interpretation of correlation.

9. Explain why zero correlations are unsatisfactory from the standpoint of the objectives of science.

SELECTED REFERENCES

Ezekial, Mordecai, and Karl A. Fox, *Methods of Correlation and Regression*, 3d edition. John Wiley & Sons, Inc., New York, 1959. Chapters 3 and 25.

Hyman, Herbert, *Survey Design and Analysis*. The Free Press, Glencoe, Illinois, 1955. Chapters 5 and 6.

Kendall, Maurice, *Rank Correlation Methods*. Charles Griffin, London, 1948.

Robinson, William S., "Ecological Correlations and the Behavior of Individuals," *American Sociological Review*, XV, 1950. Pages 351–357.

Simon, Herbert A., *Models of Man, Social and Rational*. John Wiley & Sons, Inc., New York, 1957. Chapters 1-3

Snedecor, George, *Statistical Methods*, 5th edition. The Iowa State College Press, Ames, Iowa, 1956. Chapters 6 and 7.

couple in a large city, he may find such a program impracticable to execute because of limited financial resources. He will then be content with a carefully selected sample. But this is not an unusual restriction. Research workers in general expect to arrive at valid generalizations on the basis of sample materials, presumed to be representative of some wider domain. Thus, the anthropologist interrogates a few native informants in order to reconstruct the entire pattern of a culture; the U.S. Children's Bureau relies on a sample of juvenile courts in order to establish national trends in delinquency.

Moreover, there are times when sampling is not only practically advantageous, but an almost inescapable necessity. When we destroy an object in the very act of measuring one of its characteristics, we must either sample only a few or destroy them all. This goes by the name of *destructive sampling*. For example, it is impossible to measure the life span of an electric light bulb without ending its life; consequently, to approximate the average life span of a large lot, the manufacturer necessarily resorts to a small sample. Analogously, the physician cannot afford to sample more than a few drops of blood without causing undue discomfort in his patient.

But destructive sampling may sometimes be abandoned, if we are unwilling to tolerate the destruction of even a single item. Where human life is jeopardized, we are not willing to risk the loss of a single sample case. And even when the sampling is not harmful, social considerations inhibit us from freely carrying out experimentation on human beings. It is only when an occasional life convict or conscientious objector voluntarily submits to a possibly harmful experimental treatment that public opinion permits destructive sampling to be carried out on human beings. In the usual case, such sampling is transferred to dogs and monkeys, who serve as stand-ins for man.

Sample materials are thus the stuff of which scientific generalizations are made, but the sampling method involves more than mere recognition of that principle. Modern sampling practice is distinguished by (1) its emphasis on the well-defined universe, (2) the random selection of cases, and (3) the estimate of the reliability of the sample statistic — that is, how closely it probably conforms to the unknown parameter. These concepts form the basis of the discussions in this and the following chapter.

The Statistical Universe (Population)

Definition of Universe. In colloquial speech, the term *universe* suggests the entire Creation. But in statistical language, it refers merely to a totality of values possessed by elements having a well-defined common characteristic which determines membership in the set. Such sets are also termed statistical *populations*. We may cite as universes: the incomes

Moreover, fresh samples of experience continually stimulate us to reexamine our previous inferences. Thus, the success of one woman in Congress will cast doubt on the previously held opinion that women are politically incompetent; the grade records of Negroes in northern schools will force a revision of the stereotyped belief that Negroes possess an inferior mentality. Thus, we modify our premises in response to the unfolding sample evidence. In the language of statistics, we afford our initial hypotheses an opportunity to be revised or nullified.

While every man thus reasons informally from part to whole, statistical inference is much more rigorous than such mere folk practice. Sampling has come to be a body of technical procedures which must be deliberately *applied and strictly adhered to if its goals are to be fully realized.* Thus, the statistical universe must be well defined, and the sample must be properly drawn — two basic operations which are much more involved than appears on the surface. In fact, the sampling system, as it has evolved in the twentieth century, is essentially a technical accounting device to rationalize the procurement of information. By this criterion of efficiency, it is always preferred to complete enumeration when it yields data of requisite accuracy, since such a procedure is obviously more economical in a busy world of limited resources.

Further, the logic of sampling is reinforced by the generally accepted view of nature as being orderly and predictable. The discipline of statistics has not always enjoyed such a favorable climate of opinion, especially when applied to human behavior. In the mid-nineteenth century, Quételet and Buckle were denounced as materialists because they drew inferences from statistical records on the recurrences of crimes, suicides, and other human actions that ran counter to the then current doctrine of free will. It is therefore no accident that the sampling method has been cultivated most intensively and that the sampling mentality has flourished in the scientifically-minded, efficiency-oriented rational culture of the modern Western world.

Advantages of Sampling. The collection of sample data naturally requires less time and effort than does the compilation of complete data. Hence, surveyors of American public opinion universally avail themselves of samples of respondents, since results based on time-consuming total enumerations would be obsolete before they could be tabulated and published. Similarly, since 1940, sampling has been extensively employed by the U.S. Bureau of the Census in its decennial enumerations to provide more promptly detailed descriptions of various characteristics of the American population.

Since sampling is also less costly, it may be quite feasible when the financial burden of full coverage is prohibitive. Although a free-lance sociologist, for example, may believe it desirable to interview every divorced verse contains a countable number of elements. It may be relatively small, as, for example, all students enrolled in a particular college in a given year; or it may be relatively large, as all college enrollees in the United States. But the infinite universe consists of an endless number of elements, such as an unlimited number of penny tosses or other experimental trials. It is thus purely conceptual, and may even seem metaphysical to the finite mind. And yet it is often heuristically postulated in statistical inference. For one reason, an infinite population permits the reliability of the sample findings to be more simply evaluated by formula than does a finite population. Consequently, we resort to the assumption of an infinite population whenever the size of a finite population is large enough to justify it. Additionally, it may be invoked when it is not reasonable to limit the size of the universe at all, as in the case of infinitely repeatable experimental trials. The laboratory dog may die after an experimental injection; but the scientist does not restrict his generalization to the dead dog. His ultimate interest lies in the potentially endless succession of experimental trials ideally performed under identical conditions. Onto this infinitely large universe he fastens his generalization. Analogously, but at considerably greater risk, the social scientist may conceptually extend his findings in one or a few cases to the hypothetical infinite class of presumably identical events. One family, one community, one culture, one bureaucracy — each in its turn serves as a prototype to which all future occurrences of the same general class are presumed to conform.

(3) The universe which is actually sampled (the *sampled universe*) will not always coincide with the universe on which our sights are fixed (the *target universe*). The target universe represents ideally the territory we intend our generalizations to cover — the domain to which we eventually apply our sampling knowledge. Our ultimate interest may lie, for instance, in the patterns of adjustment of all married students on campuses in the United States, but for practical reasons it may be necessary to restrict the sampling to the available couples on a particular campus, and these couples then become the sampled universe. When a mailed questionnaire is returned by less than 100 per cent of the sample, as is usually the case, we may conceive of the target universe as the complete mailing list from which the sample was selected, and the sampled universe as all persons on that list who would theoretically return that questionnaire if given an opportunity to do so. Thus, the concept *target universe* may be applied broadly to an idealized extension of the sampled universe, or narrowly to a universe in which a fraction of the units are for one reason or another inaccessible to measurement, such as refusals or not-at-homes.

Strictly speaking, statistical inference should be rigorously limited to the sampled universe; and yet the social analyst can hardly refrain from speculating about his target universe. No research study in the social of all American families in 1960, the opinions of all college students on the subject of war, or the social-status ratings of all residents in Yankee City. From these examples, we discern that the statistical universe may be conveniently conceived of as being dualistic in nature: (a) one dimension consisting of the units (e.g., families) which are actually sampled and which are therefore called *sampling units*; (b) the other dimension being the *sampling trait* (e.g., income) possessed by the sampling units, which is *subsequently manipulated statistically*.

It is the sampling units that are physically selected: families with incomes, farms with acres, students with opinions, workers with occupations. Further, the well-defined *criterion which determines eligibility* for such selection here applies to the sampling units rather than to the traits. Thus, a population of families consists of all human groups that satisfy the working definition of a family.

On the other hand, it is the variable properties of the sampling units that command our ultimate interest. We are seldom if ever interested in families *qua* families; rather our interest will lie in one or more of their relevant traits: income, nationality, size, social status, religion, and so on. It is variables such as these that are subjected to statistical measurement after the sampling units have been drawn and their characteristics determined.

When, for any reason, it is unnecessary to identify explicitly both aspects of the universe, we may quite properly employ elliptical statements such as "the population of college students" or "the universe of attitudes." But such abbreviated statements omit the sampling trait of attitude (of the college students) in the first instance, and the sampling unit, the college student (whose attitudes are polled), in the second instance. Although terminology varies among statisticians, we here conceive of the "universe" as comprising both the sampling unit and the corresponding sampling trait.

Universe Classification. Statistical universes differ in various ways and may therefore be classified according to various criteria. Three of the more important classifications are here briefly set forth to amplify the sampling concept and refine its implementation:

(1) The universe will be *qualitative* or *quantitative* according to whether the traits of which it is composed are attributes or variates. This familiar distinction implies that the statistical description of the universe will take the form either of arithmetic averages or frequency counts and percentages. For example, the incomes of American families may be represented by the mean, whereas the occupations of American workers can only be grouped and counted and expressed as percentages of the total.

(2) The universe may be *finite or infinite*, depending on whether the sampling units are finite or infinite in supply. By definition, a finite uni-

using persons enrolled in English Composition, as this subject is required of all. Again, reasonable enough on first thought, this alternative will seem less satisfactory when we realize that only underclassmen are enrolled; mature upperclassmen would have no opportunity to be included in the sample.

Many other possibilities will readily suggest themselves in the search for representative coverage. Thus, we might consider canvassing the men's dormitories. But this scheme rather glaringly omits women, and so would have to be modified to include women's residence halls as well. Even with this modification, the plan is rather obviously unsound: it makes no provision for students housed in fraternities or for those living at home. Apparently, other plans will have to be tried until an adequate one is developed.

The general shortcoming of the aforementioned alternatives will be recognized by even the casual reader: they do not afford each sampling unit an equal opportunity of being selected. Ideally, of course, our sampling procedure should give to each unit in the population such an opportunity. Methods which meet that criterion are named *random sampling procedures;* methods which offer no such assurance may be conveniently labeled *non-random procedures.* When such equal opportunity of selection is not provided for by the sampling technique, it will generally be impossible to vouch for the representativeness (reliability) of the sample, and consequently it will be impossible to generalize confidently from the sample back to the universe. The fulfillment of the purpose of sampling is jeopardized.

Non-random procedures are therefore seldom considered to be ideal; nevertheless, they are often justifiably resorted to in social research because of practical necessity. Beginning with the least useful, several are presented here in ascending order of utility so that the student will be familiar with their respective merits and shortcomings.

Non-random Procedures

Haphazard Sampling. The acceptance of whatever cases one fortuitously happens to encounter, without any consideration whatsoever for their degree of representativeness, may be termed *haphazard* sampling.* This practice is exemplified by the old-fashioned straw vote in which citizens are accosted in the street to ascertain their voting intentions. From these straws in the wind, the election "forecast" is made. The obvious objection to such casual procedure is that the man-in-the-street simply is not representative of the total electorate. Similarly, in a survey of student opinion on the propriety of final examinations, those

* Synonymous concepts that appear in statistical writing include *accidental sampling* and *convenience sampling.*

sciences would ever be made if its findings could not be imaginatively extrapolated beyond the limited universe from which the sample has been derived. Such speculations are both justifiable and desirable, provided that their tentative nature is recognized and understood. A sampling study of mental health and social class in a New England community is of significance primarily for the light that it sheds on the relation between social structure and personality deterioration in other American communities. And so the investigator quite naturally probes his sample materials for their wider generality and makes the most of his costly data. In projecting his findings onto the vaguely defined target universe he necessarily proceeds without benefit of strict reliability procedures. Yet he may be, and usually is, engaged in fruitful and necessary scientific activity. Nevertheless such a liberal statistical morality, which is practiced by even the most thoughtful and productive social scientists, is by no means license for irresponsible and sloppy statistical generalizations. It should be permitted *only to experienced investigators.*

Sampling Procedures

Problems of Sampling. The process of sampling is in its scientific sense a technical operation which must be conducted according to standard prescriptions in order to secure all of its benefits. In fact, costly social investigations have sometimes been severely blemished because the sampling tactics were crude and inadequate. Because of the admitted difficulties in sampling human populations, the discussion of the theory and practice of sampling must always occupy an important place in the domain of social statistics.

Consider, by way of illustration, the task of sampling a set of college students — a common assignment for majors in journalism or sociology — with a view to generalizing about the entire student body. Disregarding momentarily the kind of data sought — attitude toward communism, number of dates per week — how may we obtain a sample which will do justice to the student body? This is no simple task. The sampler might give way to the first impulse to take all persons enrolled in elementary sociology — a rather attractive possibility since such classes include a wide variety of students and, in addition, are easily reached and manipulated for experimental or survey purposes. However, on second thought, he will realize that college courses are almost sure to exert some selective influence among students. Thus, sociology is likely to attract persons whose primary interest is in social issues, but may hold no appeal for those principally interested in the physical world. In restricting the sampling to sociology students, there is therefore a danger of excluding certain types of individuals who are negatively selected by this subject. To forestall this outcome, we might propose nighthawks whom we happen to find breakfasting in the coffee shop at 11 A.M. will be something less than a fair cross-section of the entire student body. Such crude chunks of data are "samples" only in the loosest sense of that term. When seriously used, they constitute an unflattering reflection on the sophistication of those who resort to, and accept, such data.

Availability Sampling Although fully aware of the limitations of nonrandom sampling, the experienced social scientist will sooner or later realize that some form of it is often the only alternative to abandoning the inquiry. In many instances, he may be required therefore to seize whatever opportunity is available. The Kinsey survey of sexual behavior in the United States male popnlation was severely criticized because it was based to a large extent on data provided by solicited subjects who made themselves available. It was plausibly contended that persons who volunteer to provide information about their sexual behavior are likely to differ significantly from persons who decline to be interviewed. But the investigators considered the acceptance of volunteer respondents as a pragmatic solution, rather than the preferred procedure. It was their reasonable assumption that many random selectees could not be induced to relata their sex histories, and that any possibility of obtaining a completely random sample was precluded.

For similar reasons, other types of sociological studies must oftan rely on available opportunities. For example, it may be impossible for the behavioral scientist to sample all workers in a given industry, but yet possible to observe workers in a local plant; impossible to poll a group selected from all school children in the state, but possible to poll those attending various local schools. Social psychologists are likely to find little humor in the sarcastic remark that their broad "universal" generalizations are founded on experiments on college sophomores. Above all, they would prefer to study all kinds of individuals — old as well as young, non-college as well as college — in formulating the laws of social learning. But it is often a question of sampling the available students or otherwise abandoning the project. Hence, they must convert a captive audience of students into a "valid" sample. Similarly, research on delinquency is almost always based on the available delinquents — boys on probation, boys in court, or boys in the industrial school. Regardless of how desirable it might be to study a group selected from all delinquents, the unapprehended as well as the apprehended, such comprehensive sampling is for obvious reasons out of the question.

It would therefore be pedantic to deny the uses of available opportunities, even though they do not yield ideal data. Social scientists, like everyone else, must often content themselves with compromises. Notwithstanding their shortcomings, availability samples do yield signifiin 25. How is this to be interpreted? The layman would reply: "Since the human sampler cannot influence the drawing, eacb guest has the same chance." But the statistical interpretation lies in the very definition of probability, which applies to mass phenomena or conceptually repeatable trials. In this instance, "same probability" refers to the expectation that each name would be drawn an equal-number of times if identical drawings were conducted indefinitely. Of course, an empirical demonstration of innumerable drawings is not undertaken; nevertbeless, it is tacitly assumed that the a priori expectation of equal occurrences would eventually be confirmed. This principle is intuitively recognized hy the losing guests wbo confidently console themselves witb tbe bromide, "better luck next time."

Of course, equal probahility of selection cannot he assumed if some cases are less accessible than others or if the selection mechanism is functioning imperfectly. Either of these related contingencies will defeat the aim of random sampling and result in *biased sampling*. Thus, in selecting a sample of 10 slips from a receptacle containing 200, randomness would be precluded if tbe slips were unequal in size, shape, or weigbt; or if they were carelessly mixed so that the last names dropped into tbe hat would have the best opportunity of being drawn. But the outward appearance of the extracted sample would never reveal sucb procedural flaws, or biasing factors. The composition of a non-random sample is visibly no different from that of a sample of the same size selected by random procedures; tbe sampling operation leaves no telltale mark. Tbus, if we were to come upon two different scatters of ten coins eacb strewn on two tables, the one all heads and the otber showing six heads and four tails, no amount of visual inspection would tell us wbether eitber or botb of tbem bad been carefully laid down or whetber they bad been tossed at random. Nevertheless, the observer will intuitively conclude that the first had probahly heen laid down and the second had been arbitrarily tossed. Why? Because the probahility of 10 heads together is so small, and the division of 6 and 4 much greater. In fact, the probabilities compare as 1 to 210. Still, he can never be certain. Similarly, in the case of our hostess, the name on the winning slip can never betray whether the drawing was honest or biased.

However, a suspicion of biased sampling will be aroused by the subsequent discovery of a marked discrepancy hetween the sample value and the true value of the universe, when and if the latter becomes known. The now classic example of such a melancholy outcome is the notorious *Literary Digest* Presidential pre-election poll of 1936. While Roosevelt obtained approximately 60 per cent of the popular vote in the actual election, his percentage in the *Literary Digest* sample of over 2,000,000 respondents was only 40 per cent — a difference of 20 percentage points! This discrepancy in such a large sample was symptomatic of a gross ject to human bias and error. And even when the universe is in plain sight, it is probable that the observer will misjudge to some extent the representativeness of a sample. To demonstrate such bias, the English statistician Yates once requested 12 persons to select three samples of 20 items each from a collection of 1,200 stones, each sample to represent as accurately as possible the size distribution of that experimental universe. Although the observers were free to view the collection at their leisure, still there was a consistent tendency to exaggerate the average size of the stones and to minimize their variation. In short, there was a constant error in judgment. We cannot be sure that analogous errors did not bias the "Fa" sample. Because of the practical certainty of human bias, judgment sampling must therefore be applied with great caution. But whatever the limitations of non-random samples may be, the techniques of descriptive statistics must obviously be competently applied in order to assure their maximum utility.

Random Sampling Procedures

Definition of Random Sampling. If available samples are fallible, and expert judgment is not to be trusted, then what are the factors that should determine the composition of the sample? The answer is: chance factors. However, in view of the low regard in which chance factors are usually held as guides to action, it is surprising that we should so willingly lay aside our cumulated knowledge and experience and go to the other extreme, permitting blind chance to determine the choice of the sample. In most human situations, we wish to eliminate chance, since it disturbs our predictions. Nevertheless, the ideal sampling procedure is one in which the drawings are affected by impartial chance factors alone, with the result that one item in the universe is as likely to be included in the sample as another. No item is accorded a preferential advantage. In fact, *random sampling* is defined as a procedure that provides an equal opportunity of selection to each unit in the population. We eliminate personal bias by elimination of the person, and thereby permit only the play of impersonal chance forces.

There is nothing esoteric about random sampling; most persons are familiar with a few homemade routines that assure an impartial randomized selection. The thoughtful hostess wishes every guest to have an equal opportunity to win the door prize, regardless of wealth or friendship; the contest is to be perfectly democratic. Accordingly, everyone is requested to write his or her name on a blank slip and drop it into the hat. After the last arrival, the slips are thoroughly mixed, and one name is drawn and declared the winner. Although only *one* person can win the door prize, we still say that the probability of winning was identical for all. Among 25 guests, the probability of winning would be 1

For finite universes and samples of any size, this probability may be expressed in terms of the now familiar combinatorial formula:

$$Pr\text{ (Given Sample)} = \frac{1}{C_n^N}$$

Applying these formulas to the illustrative data given above, we have

$$C_n^N\text{ (Total Possible Samples)} = \frac{5 \times 4}{2 \times 1} = 10$$

$$Pr\text{ (Given Sample)} = \frac{1}{10}$$

which agrees with the previous results.

Sampling by Random Digits. A rudimentary technique for carrying out simple random sampling has already been set forth: (1) represent units on slips; (2) thoroughly scramble; and (3) draw the required number of slips. However, the drawing need not be laboriously carried out in this manual fashion. It will usually be inefficient to do so, especially when the population is large. It is much more practicable to substitute a table of randomly ordered digits for the shuffled numbered slips — a common procedure of research workers. This technique necessarily requires that we number serially all of the units in the population from 1 through N and then draw from the corresponding table of random digits as many different numbers (combinations of digits) as there are cases to be included in the sample. The cases whose serial numbers correspond to those drawn from the table constitute the sample. Table IV of the Appendix presents such a list of random numbers.

Other Random Sampling Procedures

Simple rondom sampling is the most primitive and unrestricted selection procedure and most clearly exposes the essential operation of randomness. Being free of procedural modifications which are often made necessary by practical circumstances, it is conceptually the simplest of all sampling routines and is therefore so labeled. Since it is random sampling in its most uncomplicated form, it serves as a standard of sampling efficiency against which other types are compared and evaluated. These other types are made necessary by the fact that simple random sampling in its pure form can almost never be employed in large-scale social research. It is far too impractical and costly, and often even impossible. Nevertheless, it is essential that the student clearly comprehend its basic characteristics in order to be able to recognize and appreciate the degree to which the alternative methods depart from this standard model. Three of the most prevalent alternative types are here set forth as a brief defect in the sampling procedure, which, upon later review and analysis, statisticians easily established. Although such a discrepancy could have occurred by chance, it would have been extremely unlikely. Hence the deduction that the method of sampling was biased. But such appraisals are always retrospective. In advance of the sampling, one can only provide for adequate machinery which will be reasonably certain to yield a random selection.

Simple Random Sampling. If called upon to devise a do-it-yourself sampling technique, the inexperienced layman, like our hostess, would probably procure as many slips as there are items in the population to be sampled. Slips of paper are more easily shuffled than people. We may surmise that he would then number these slips consecutively from 1 through N, corresponding to the numbered units of the universe. Next, he would place the slips in a suitable receptacle and mix them until the set was thoroughly scrambled. Finally, he would reach in and take out as many cases as desired. If he wished a sample of 10 cases, he would take out 10 different slips. Statisticians recognize this as the simplest type of random sampling and have therefore dubbed it *simple random sampling*.

They do not casually describe it as "reaching in and drawing out n different items." Rather they define it as that procedure in which every distinct sample of n items has an equal probability of selection from a finite population of N items. This definition expresses the long run consequence of "reaching in and taking out n different items." For, if that procedure were applied indefinitely to a given population (restoring the entire sample after each trial), each different sample would tend to reappear an equal number of times. In its procedural aspect, simple random sampling is the apparatus that guarantees the fulfillment of this criterion of equal probability. In its substantive aspect, it is the very criterion itself.

To unfold further the meaning of simple random sampling, let us consider the number of ways in which samples of 2 items can be selected from a miniature population of 5, whose members we shall designate: a, b, c, d, e. By manipulation, we discover that there are 10 different possible combinations, or samples:

ab
ac
ad
ae
bc
bd
be
cd
ce
de

Hence, the probability of each combination in simple random sampling must be 1 in 10; each sample is expected to occur on the average once in every 10 trials.

from each of the hroad geographical regions of the United States, it being plausibly assumed that public opinion on many issues varies from one sector to another: the Midwest, for example, is internationally more isolationist than the East. Similarly, subsamples are commonly drawn from various age and economic levels, as these factors are also known to exert an influence on the content and intensity of public opinion. Older persons are generally more conservative than younger; political opinions tend to parallel economic interests. On the other hand, national opinion polls would never stratify the population hy hair color, since there is probably no genuine correlation between hair color and political opinion. In general, stratification serves no purpose when the stratifying factor is uncorrelated with the sampling trait being measured.

Comparison of Strata. While the principal justification of stratified sampling is equal accuracy with a smaller sample, the comparative data which are its natural hy-product provide an additional inducement for using it. Thus, the decision to stratify hy occupation in a survey of opinion on labor unions may be prompted as much by the wish to compare the characteristics of the various occupational classes as hy the need to economize on sample size. The differences among occupational categories (Tahle 11.1) on the question "Are you in favor of lahor unions?" may, in fact, be even more pertinent and revealing than the over-all weighted average of 67 per cent, which necessarily conceals such differences. Since strata are often treated individually, it has been suggested that the term "domains of study" be applied to strata when they are being analyzed in this segregated manner.

Table 11.1 *Attitude Toward Labor Unions by Occupation, Percentage Distribution, U.S.*

Occupation	Favorable	Unfavorable	Total
Farmers	52%	48%	100%
Businessmen ...	66	34	100
White Collar....	69	31	100
Professional.....	77	23	100
Skilled. .. .	75	25	100
Unskilled.... ..	71	29	100
Total	67%	33%	100%

Source: American Institute of Public Opinion, May, 1942.

introduction to the subject: (1) *stratified*, (2) *cluster*, and (3) *interval* sampling. These sampling devices themselves are largely an outgrowth of the sheer practical problems that have arisen in the sample surveys of large human aggregates, which accounts for their wide currency in social science.

Stratified Sampling. "Do the citizens of Brownville favor racial integration of schools?" No competent surveyor of public opinion would attempt to answer that question without canvassing both white and Negro residents of the community. Public sentiment on this issue would not be reliably portrayed by a sample that sligbted eitber group, since their opinions are so divergent. Yet sucb an imbalance between groups might occur under *simple random sampling*, unless the sample were made large enough to forestall that eventuality. A more economical alternative would be to sample each subgroup separately and combine the results, and thereby avoid a costly inflation of sample size. Such an operation, which first separates the entire population into relevant *strata* before randomly drawing the sample, is known as *stratified sampling*.

From a procedural standpoint, stratified sampling therefore consiets of the following etages: (1) division of the total univeree into eubclasses, or *strata;* (2) the selection of a random sample from eaeh stratum; and (3) the eonsolidation of the subsample statistics into a comhined statistic weighted for size of etrata. In this eontext, the term *stratification* does not, of course, eonnote a hierarchy, as the ranks of an army, or the geological seams of the earth crust; rather it signifies the categories of a statistical variahle, such as race, sex, or religion, into which the total population ie eonveniently divided.

Nor does stratification imply a relaxation of the requirement of randomized selection, althougb that inference has been eometimes mistakenly drawn. This misconception may possibly reflect a failure to distinguish clearly between stratified random sampling and so-called *quota sampling*, which was formerly widely used in opinion polling. In quota sampling, quotas are pre-assigned to strata, but the final selection of cases is left to the discretion of the interviewer. However, if the benefits of random sampling are to be attained, subsamples from strata must be randomly chosen. The resort to convenience or judgment sampling is no more warranted within a given stratum than it is in the whole, unstratified population, and is almost certain to lead to biased results.

Since stratified sampling is more complex than simple random sampling, we may rightly ask what are its compensating advantages. Briefly put, it is a labor-saving device for securing equivalent accuracy with fewer cases than is likely under simple random sampling. It is essentially for this reason that national public opinion polls universally resort to some form of stratified sampling in order to keep the size of the sample down to manageahle proportions. Thus, subsamples are usually selected

However, the execution of any social survey is not fulfilled by the mere drawing of the sampling units. We still must make the personal contact with the selected elementary units to obtain the information which was the objective of the study in the first place. This problem of personal contact develops into a major obstacle in simple random sampling, which may yield a sample whose elements are widely dispersed over a wide area, requiring a prohibitive expenditure of time and energy to reach. It is in such a plan for personal interviews that cluster sampling displays another technical advantage. Under this procedure, the elementary units are territorially concentrated, and are therefore more easily accessible, with less wastage in transportation. This is shown in Figure 11.1.

Simple Random Sample

Cluster Sample

FIGURE 11.1 *Schematic Diagram, Simple Random and Cluster Sampling*

But the physical and mechanical convenience of cluster sampling is purchased at a price in quality. It generally lacks, to a degree, the very characteristic which is the objective of good sampling: typicality and representativeness. Its reduced representativeness is due to the fact that the elementary sampling units within clusters, particularly human clusters, are likely to be closely similar in regard to their social characteristics; consequently, sampling within these clusters understates the dispersion and provides unnecessary duplication. For example, the residents of a given city block are likely to belong to the same socio-economic class, and *therefore hold the same political* and *social opinions. By* the same token, they are not likely to be representative of any diversity of public opinion in the large aggregate of unsampled neighborhoods; a cluster sample is thus *less* likely to represent the variety of opinions than a simple random sample of the same size scattered over the community. It is this lack of heterogeneity within clusters which reduces the comparative effectiveness of cluster samples.

This liability notwithstanding, cluster sampling is being increasingly employed in social research. For, after all, statistical work — like every other human endeavor — is a compromise between the ideal and the practical, and in any case can attain only an approximation of "truth."

Cluster Sampling. In the case of the drawing for the door prize, it was obviously necessary to have a complete list of eligible names, from which the winning one was randomly drawn. For, "equal opportunity to be included" implies not only a mechanically reliable selection procedure, but also a complete list, so that every name is accessible to the draw. It would similarly be necessary to have a complete source list of sampling units if a simple random sample were drawn from any other statistical universe, such as the inhabitants of a community, an army, or a university student body. But such social universes differ from the more or less trivially small, easily manipulated guest list of a banquet. When the universe is vast, and extends over a wide area, the compilation of such an indispensable list is a laborious undertaking; and even when available (e.g., a city directory), it is usually not up-to-date on the date of issue because of the mutability of the population.

If, therefore, an adequate list of elementary sampling units is not available, we may turn instead to more or less permanent groupings into which the population is naturally divided and which can be conveniently listed. Human beings are, of course, usually found in prevailingly standard groups: they are clustered geographically by states, counties, municipalities, neighborhoods, blocks, precincts, and dwelling units; people work together in factories, offices, and stores; they are organized in innumerable clubs, lodges, schools, and miscellaneous associations. It is these groups, or *clusters*, which may be serially utilized as sampling units, through which we reach the ultimate elementary unit (e.g., the person or household) which is the objective of our survey. Such a procedure is therefore termed *cluster sampling*. We thus finally reach the elementary unit through a shorter or longer chain of samplings of the more easily listed clusters. Since sampling is carried out in successive stages before reaching our destination, this type of sampling is also called *multi-stage sampling*. When the clusters at any stage consist of territorial units, we may describe that stage as *area sampling*.

Let us suppose that we wish to survey the occupational ambitions of high school pupils, ages 15–17, in Chicago. It would be exceedingly laborious and financially prohibitive, as well as hazardous in accuracy, to compile a list of all specified pupils in that metropolis. However, a permanent list of high schools is easy to obtain. These schools could be randomly sampled, and a list of students in the desired age categories would then be obtained from the comparatively few high schools in the sample. From that list, the sample of students would be finally drawn. Similarly, if the households of the city were to be surveyed, a sample of city blocks, then of dwelling units, and finally of households would be drawn. Even though such a multi-stage procedure would still require a source list at each level, the lists would be smaller and more current. In such economy lies the first advantage of the cluster approach.

presumed to be randomly ordered, simple random and interval sampling will yield identically accurate results. Such a presumption is often reasonable when items have been alphabetically listed, since there is usually no correlation between the alphabetical order of names and the traits which the named objects possess. Thus, we have previously seen that the alphabetical listing of large American cities orders the respective suicide rates in a sequence which is seemingly purely random. Similarly, an alphabetical listing of students may be expected to result in a sequence of grade averages that is wholly random and, therefore, free of trends and cycles. In such cases, it makes no ultimate difference whether we select random digits or draw every *k*th unit, except that interval sampling is usually more simple to execute. The long-run results would be virtually identical for any given size sample.

There is, however, one notable circumstance that constitutes a special hazard for interval sampling and may easily lead to erroneous conclusions. When the universe values form a *cyclical progression*, the sampling interval may coincide with the phase of the cycle, causing interval sampling to yield an unrepresentative set of identical values. Let us consider a fictitious sequence whose phase is four: 1, 2, 3, 2; 1, 2, 3, 2; 1, 2, 3, 2. Now, if the sampling interval is set equal to 4, any sample will necessarily consist of a set of identical values. It will consist of all 1's, all 2's, or all 3's. Such samples will in no way do justice to the variation in the universe.

This type of pitfall is illustrated in a sampling study of June issues of the Sunday *New York Times*, 1932–1942, which disclosed that only Protestant marriages were featured on the society page of the sampled issues.* From this finding, the conclusion was drawn that the upper-upper social class of New York City was preponderantly Protestant in religious background. But this inference was immediately challenged on the ground that Jewish marriages happen for ceremonial reasons not to be performed in June, and therefore had no opportunity to appear in the issues of the *Times* which were sampled. A check sampling, undisturbed by daily and monthly cycles, revealed that Jewish marriages were in fact regularly featured by the society editors during the appropriate seasons. By that criterion, Jews were proportionately represented in the upper social strata. In this instance, the sample interval led to an overrepresentation of Protestants, an error compounded by the unfortunate judgmental selection of June as a point of origin.

But interval sampling also carries its intrinsic advantages. In fact, when the numerical values form an *arithmetic progression*, interval sampling will be even more effective than simple random sampling. For, in

* David and Mary Hatch, "Criteria of Social Status as Derived from Marriage Announcements in the *New York Times*," *American Sociological Review*, XII, 1947, pp. 396–403; and W. J. Cahnman, "A Note on Marriage Announcements in the *New York Times*," *American Sociological Review*, XIII, 1948, pp. 96–97.

Interval (Systematic) Sampling. Whenever the sampling units are arranged in some kind of natural sequence, like consecutive admissions to a hospital or library books in the card catalogue, it may be economical and even preferable to obtain a sample hy taking cases at a fixed interval. Such a procedure is termed *interval sampling*, or more commonly but less aptly, *systematic sampling*. The selection of every tenth name from the telephone directory, after a random start among the first ten names, illustrates the process of intervnl sampling. Such a procedure has obvious utility for the social scientist wbo frequently has occasion to study a series of events such as a file of newspapers, the characteristics of dwelling units in a given ecological area, the cases on docket in a criminal court, or a card catalogue of welfare case records.

To establish the widtb of the sampling interval (k) in any given problem, we merely find the ratio of population size (N) to desired *sample* size (n):

$$k = \frac{N}{n}$$

Thus, if tbe sample is to contain 5 per cent of the universe, or 1 out of every 20 cases, the sampling interval ohviously would be 20; and we would dmw every 20tb item, randomly starting with any number within the first interval of 20. Such calculation presupposes, of course, tbat sample size has been fixed in advance of the sampling. But if n bas not been set, and an arbitrary interval is employed, it is still necessary to pass through the entire sequence, even though we may seem to have an ample number of items after we have proceeded only part of tbe way. If we discontinued the drawings before completing the entire circuit, we would deprive tbe units in the omitted segment of the opportunity of being chosen and thereby destroy the randomness of the opemtion. For example, by skipping names L to Z in an alphabetical listing, we would *almost certainly produce a biased sample.*

Since each interval contributes one and only one item to each sample, it follows that there can be no more *different* samples *than* there are items within the interval. Thus, if the sampling interval is equal to 10, there can be only 10 possihle samples, regardless of the size of the universe, be it 1,000 or 1,000,000. But the number of different samples would be almost incalculable in simple unrestricted random sampling. This restriction in the number of possible samples serves to distinguish interval sampling from simple random sampling, since the latter furnishes an equal opportunity of selection to every distinct combinatorial sample of n items.

In spite of this severe limitation, interval sampling will often produce results that compare favorahly in representativeness to those yielded by *simple random sampling.* In particular, whenever the values may be

Analogously, the effectiveness of cluster sampling will be enhanced by expertly composing and recomposing clusters in advance of the sampling so that each cluster is as representative of the entire population as possible. Insofar as that effort can achieve success, given the heterogeneity of the grand universe itself, the respective clusters will tend to resemble one another, while there will be considerable statistical variation within each cluster. Thus, in cluster sampling we invert the specifications of stratified sampling: instead of homogenizing strata we diversify the elements within clusters. We may, for example, combine precincts into larger geographic districts in order to increase the diversification within clusters and thereby raise their representativeness of the entire electorate and fulfill their function as samples.

In sum, no sampling technique is completely automatic; all involve subject-matter decisions. Hence, first-hand practical experience with the concrete subject matter contributes quite as much to fulfillment of a sampling project as dexterity in the mechanical routines of applied statistics.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Universe
   Sample
   Infinite Population
   Finite Population
   Sampled Population
   Target Population
   Availability Sample
   Haphazard Sample
   Judgment Sample
   Random Sample
   Simple Random Sample
   Stratified Sample
   Cluster Sample
   Interval Sample

2. State difficulties that might be encountered in defining the populations:
   Families
   Dwelling Units
   Farms
   Households
   City Blocks
   Broken Homes
   Overcrowded Homes
   University Students
   Gainfully Employed
   Social Class

that event, the sample will necessarily distribute itself evenly over the entire range of values and thereby provide a reliable miniature of the population distribution. If, for example, we select every fifth boy from a lineup according to height, the resulting sample will necessarily be representative of the distribution of boys' heights. Analogously, if we sample every 25th dwelling unit along a metropolitan avenue after a random start, we would probably obtain an accurate cross-section of the various socioeconomic groupings on that street, since dwelling units are segregated and ordered according to social status. In this way interval sampling may supply its own stratification.

From this, it is evident that the principal advantage of interval sampling lies in the mechanical ease with which it can be applied to such natural sequences as rows of dwelling units, card files, city directories, and so on. Its special hazard is the cyclical sequence, and we must maneuver to circumvent it when that danger is thought to exist.

Interrelatedness of Random Sampling Procedures. Quite obviously, the foregoing random procedures are not mutually exclusive. They may be — and usually are — combined in a variety of ways. Cluster sampling may be used within broad strata, and interval sampling may be used within these clusters No single *sample design* is best for any given purpose. In all sampling, an attempt is made to attain the desired degree of representativeness as economically as possible, which is the guiding criterion of modern sampling design. We should recognize that the designing of a sample is a form of statistical engineering and accounting, requiring appropriate skills and knowledge. In the foregoing statement, we have merely hinted at the technical aspects of sampling, which are of course fully developed in treatises on that subject.

However, effective sampling requires much more than mere technique. If the assets of a given sampling procedure are to be fully realized, it is essential that all necessary discretionary as well as mechanical steps be expertly performed. Thus, the anticipated benefits of stratified sampling will not be attained unless the strata into which the population was judgmentally divided before the sample was drawn actually differed among themselves on the sampled trait. In the aforementioned instance, stratification by race would be profitless if Negroes and whites shared the same opinions on school integration. If on the other hand, Negroes and whites differed widely, then, statistically speaking, there would have been considerable variation between strata, but relatively little within each stratum, thereby validating the original decision to stratify by race. It is more efficient in terms of sample size to sample from two homogeneous strata than from one very mixed, heterogeneous stratum. The knowledgeable worker must anticipate the validity of the stratifying criterion before the sampling begins.

mixture of nationalities: a Polish concentration (5 per cent of the population) in letters X, Y, Z, and an Italian (5 per cent) concentration in P, R, S. What sampling procedure would you use?

18. (a) "The admissions to the hospital over a period of a year constitute a random order." Comment.
    (b) Comment similarly on the daily list of dispensary patients.
    (c) Comment on a list of prison admissions.

19. Suggest a procedure to determine whether there is a correlation between national origins and initial letter of the name.

20. Distinguish clearly between the concepts "cluster" and "stratum" in method of selection. What are the desirable characteristics of a cluster? a stratum?

21. From a class of 25 pupils, how many distinct triads (groups of three) can be formed?

22. (a) If a person were dealt a bridge hand (sample) of 13 black cards, would that prove the deck (population) consisted of all black cards?
    (b) Does it prove the drawing (deal) was biased rather than random?

23. Consider a universe of 12 elements: A to L
    (a) How many simple random samples of size 4 can be formed?
    (b) How many interval samples of size 4?
    (c) How many stratified samples of size 4, when Stratum 1 consists of elements A–F, and Stratum 2 consists of elements G–L, and 2 elements are to be randomly selected in simple manner from each stratum?

24. (a) Consult the table of random digits in the Appendix (Table IV). Write a directive for selecting by random digits a simple random sample of 100 (n) items from a universe of 13,300 (N).
    (b) In this instance, what is the sampling ratio or fraction?
    (c) How many more cases would have to be added if the sampling ratio were fixed at $\frac{1}{100}$?

SELECTED REFERENCES

Ackoff, Russell L., *The Design of Social Research.* The University of Chicago Press, Chicago, 1953. Chapter 4.

Cochran, William G., *Sampling Techniques.* John Wiley & Sons, Inc., New York, 1953.

Cochran, William G., Frederick Mosteller, and John W. Tukey, *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male.* The American Statistical Association, Washington, D.C., 1954.

Havemann, Ernest, and Patricia Salter West, *They Went to College: The College Graduate in America Today.* Harcourt, Brace and Company, New York, 1952. Appendix.

Kish, Leslie, "Selection of the Sample," in *Research Methods in the Behavioral Sciences,* edited by Leon Festinger and Daniel Katz. The Dryden Press, New York, 1953. Chapter 5.

3 Which universe could be more precisely defined: a carload of wheat or a group of all taxpayers? a shipment of electric light bulbs or all gainful workers?

4. Compare the problems of sampling a shipment of wheat with those of sampling a human population.

5. If you were to select one sample peach to take home, would you select randomly or judgmentally?

6. (a) Since random sampling gives the "deviant" case the same opportunity to be selected as the "typical," what is its justification?
   (b) How do unrepresentative cases contribute to the representativeness of the sample?
   (c) Would a sample, carefully and randomly drawn, necessarily be representative?

7. If you had to select and interview a random sample of residents (adult) in the community, what procedural difficulties would you almost certainly encounter?

8. In a community survey, what would be the disadvantages of selecting a sample of families from only one neighborhood? Under what circumstances might such sampling procedure be acceptable?

9. Explain "Stratified sampling is possible only when there is some previous information on the population."

10. How would you proceed to sample the following?
   (a) The students on a campus for occupation of fathers
   (b) Households in a city for number of gainful workers
   (c) A concert audience for socio-economic membership
   (d) A theatre audience for reaction to play
   (e) Library reference room readers for length of stay

11. Distinguish stratified random sampling and quota sampling.

12. A sample of all persons 65 and over is to be drawn in a given community. Suggest possible "standard" clusters that might be used to expedite such a plan.

13. "An area cluster sample could be re-used for successive studies of a metropolitan community over a period of one year." Comment.

14. State how you would check the assumption of the *Time* survey of college graduates that "Fa" names are randomly distributed in socio-economic and other similar groupings.

15. Along a given street a mile long, the socio-economic status of the residents rises for ¾ of a mile, and then drops sharply at the edge of the city. What sampling procedure would you use?

16. List possible non-random steps in restricted random sampling.

17. There are 1,000 households on relief in a given agency; they are alphabetically arranged in a card catalog. A social welfare administrator wishes a sample of 100 in order to determine the average amount of relief. The population is a

Statistical Inference 12

Section One

Parameter Estimation

Problems in Estimation. It should be unnecessary to repeat that a sample is useful only for the information it supplies on the characteristics of the universe. Thus, the sampling process finds its ultimate consummation in a description of the population from which the sample is drawn. But this description can be only an approximation: an average monthly expenditure of $92.75 in a random sample of college students will not correspond exactly to the average expenditure of the whole student body which the sample is designed to represent. Nor, obviously, will a second sample of $88.62 necessarily correspond more closely to the unknown parameter. If, therefore, owing to the vagaries of chance sampling, no two samples are alike and all are in error, with how much confidence can we speak of the value of the universe? Clearly, we cannot merely project the value of the sample onto the universe, and let it go at that. This procedure of adopting the sample value in lieu of the parameter, or universe value, is hedged about with as many regulations as is the procedure of sampling itself. We call this set of prescribed procedures *statistical inference*.

Since, in the practical affairs of life, it is so rare that we are able to examine a whole universe, and since sampling is so frequently the resort, the procedures of statistical inference loom quite large in the repertory of statisticians. In fact, some would actually identify the science of statistics with the problem of sampling and decision-making — an extreme emphasis which seems to the authors unrealistic, since descriptive statistics have a validity and importance in their own right.

The problems of statistical inference begin with the legitimate assumption of a discrepancy between the variable sample estimate and the uni-

McCarthy, Philip J., *Introduction to Statistical Reasoning*. McGraw-Hill Book Company, Inc., New York, 1957. Chapters 9 and 10.

Parten, Mildred B., *Surveys, Polls and Samples*. Harper & Brothers, New York, 1950. Chapters 4 and 9.

Stephan, Frederick F., "History of the Uses of Modern Sampling Procedures," *Journal of the American Statistical Association*, XLIII, 1948. Pages 12-39

Stephan, Frederick F., and Philip J. McCarthy, *Sampling Opinions: An Analysis of Survey Procedure*. John Wiley & Sons, Inc., New York, 1958. Chapters 1-3

Yates, Frank, *Sampling Methods for Censuses and Surveys*, 2d edition. Charles Griffin, London, 1953

Table 12.1.1a *Array of 100 Sample Means, n = 30*

9.9	11.6	12.2	12.7	13.1
10.0	11.6	12.2	12.7	13.2
10.3	11.7	12.2	12.8	13.2
10.3	11.7	12.2	12.8	13.2
10.7	11.8	12.3	12.8	13.3
10.8	11.8	12.4	12.8	13.3
10.8	11.8	12.4	12.9	13.3
10.9	11.8	12.4	12.9	13.3
10.9	11.9	12.4	12.9	13.4
11.0	12.0	12.5	12.9	13.5
11.0	12.0	12.5	12.9	13.5
11.1	12.0	12.5	13.0	13.5
11.1	12.0	12.6	13.0	13.6
11.1	12.0	12.6	13.0	13.7
11.3	12.1	12.6	13.0	13.8
11.4	12.1	12.6	13.1	13.9
11.5	12.1	12.6	13.1	14.0
11.5	12.1	12.6	13.1	14.0
11.5	12.2	12.6	13.1	14.2
11.6	12.2	12.6	13.1	14.3

Table 12.1.1b *Frequency Tally, Empirical Sampling Distribution, 100 Sample Means, n = 30*

Class Interval	Tally	*f*
9.0– 9.9	|	1
10.0–10.9	𝍸 |||	8
11.0–11.9	𝍸 𝍸 𝍸 𝍸	20
12.0–12.9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||	42
13.0–13.9	𝍸 𝍸 𝍸 𝍸 𝍸	25
14.0–14.9	||||	4
		100

Inspecting this lot, we see that only 4 of the 100 samples coincide exactly with the universe mean, known to be 12.4. All others contain a sampling error; in fact, every one of them would probably have shown a sampling error if we had insisted on greater decimal accuracy. We can see, however, ber of samples is called an *empirical sampling distribution.* The theoretical sampling distribution would, of course, be attained only by an infinite number of samples, since it would require that many to assure the distribution of sample means in their proper, expected proportions. The significant procedural point, however, in this type of problem, is that we can usually treat the theoretical sampling distribution of means as if it were a perfect, smooth, ideal normal curve.

Distribution of Sampling Errors. By subtracting the universe mean from each of the 100 sample means, we obtain the 100 sampling errors (Table 12.1.2a), which of course have the same curve pattern as the means themselves (Table 12.1.2b). All we have done is move the zero origin to the universe mean (12.4).

Table 12.1.2a *Array of Sampling Errors, 100 Samples, n = 30*

−2.5	−.8	−.2	+.3	+.7
2.4	.8	.2	.3	.8
2.1	.7	.2	.4	.8
2.1	.7	.2	.4	.8
1.7	.6	.1	.4	.9
1.6	.6	0	.4	.9
1.6	.6	0	.5	.9
1.5	.6	0	.5	.9
1.5	.5	0	.5	1.0
1.4	.4	+.1	.5	1.1
1.4	.4	.1	.5	1.1
1.3	.4	.1	.6	1.1
1.3	.4	.2	.6	1.2
1.3	.4	.2	.6	1.3
1.1	.3	.2	.6	1.4
1.0	.3	.2	.7	1.5
.9	.3	.2	.7	1.6
.9	.3	.2	.7	1.6
.9	.2	.2	.7	1.8
.8	.2	.2	.7	1.9

The distribution of sampling errors (Table 12.1.2b) again demonstrates that the small deviations are very numerous; the larger discrepancies are few in number. If we had drawn a single sample only, we could be practically certain that our sample mean would not have missed the true mean by more than two points, plus or minus; indeed, 68 per cent of the samples are in error by less than plus or minus one.

that the universe mean is snugly nested in the very middle of the whole array of sample means, *and that most of the sample means are compactly* huddled around the true mean of 12.4, which is practically at the 50 per cent division point. *Approximately two-thirds* of the 100 means fall between 11.4 and 13.3, or virtually within one point of the true mean. Characteristically, the sample means gravitate toward the parent mean, or the center of the universe distribution.

It now becomes evident how improbable it is that a sample mean will deviate seriously from the universe mean. Furthermore, not only do sample means cluster around the true mean, but their pattern of distribution takes on a shape unmistakably approaching the normal curve (Figure 12.1.1). If we had drawn, processed, and tabulated all * possible samples, instead of merely 100 of them, we would have a *pattern* of distribution which would display still greater conformity to the smooth normal curve. This hypothetical frequency curve of all possible sample means is labeled the *theoretical sampling distribution*, and has, as its mathematical model, the ideal normal curve. Any experimental distribution of a limited num-

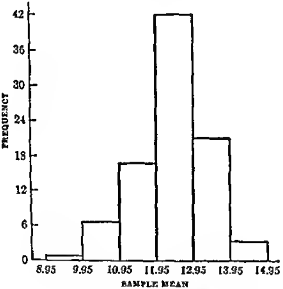


Figure 12.1.1 *Histogram of 100 Sample Means, n = 30*

* $C_n^N = C_{30}^{107} = \frac{P_{30}^{107}}{30!}$

Table 12.1.2b

Frequency Distribution of Sampling Errors

Class Interval	*f*
−3.4–(−2.5)	1
−2.4–(−1.5)	8
−1.4–(− .5)	20
− .4– .5	42
.6– 1.5	25
1.6– 2.5	4
	100

The presence of sampling errors obviously complicates the task of drawing an inference about the parameter; but tbe neat pattern of these errors renders the uncertainties of such inference less formidable. Our dissection of the empirical sampling distribution has given us considerable reassurance of the reasonableness of tbe behavior of samples, whicb will permit us to estimate tbe degree of confidence of whicb the single sample is wortby.

Sample Reliability. We must now take up the tecbnique whicb will make appropriate allowance for the ever present sampling error, and tbereby enable us to make reliable estimates of the parameter. For, unless we can rely with a certain degree of confidence on our inferences, there is no point in making them at all. The entire purpose of sampling is to substitute a sample of given reliability for the prohibitively complete enumeration of tbe unknown universe.

But we cannot measure the sampling error of our single sample directly since we are never given the true mean from which to compute it. That would suppose that we already have what we have set out to find in the first place. The single sample must ultimately therefore provide two kinds of information, as can be deduced from the foregoing analysis: (1) an estimate of the parameter, and (2) the degree of confidence we may place in that estimate, or the *reliability of the estimate* — in our example, the reliability of the mean.

The procedures for measuring the reliability of an estimate are numerous, since they differ according to the sample statistic (for example, whether mean or percentage) and according to the make-up of the sample (whether large or small, simple or stratified). Here, we present only two of the simplest techniques for measuring sample reliability; these apply to means and percentages of *large random samples*, respectively. They will suffice for our purposes, since other techniques produce results which in principle carry the same interpretatinn as the method here unfolded.

Interval Estimate of the Mean. Let us suppose that, in taking a single sample ($n = 30$) from the 107 suicide rates, we obtain a mean of 13.5. In such a realistic situation, we do not know tbe value of the true mean, and

Computation of the Standard Error. Both of the foregoing insights are in accord with statistical theory, and are actually explicit in the formula for the standard error of the mean, which contains n and σ as its basic terms. *Without further proof, we give the formula:*

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{12.1.1}$$

where $\sigma_{\bar{X}}$ = standard error of the mean
σ = standard deviation of the universe
n = sample size

Of course, we do not have the standard deviation of the universe, as required by the formula; all we have is the standard deviation, s,* of the sample. We will therefore substitute it for the unknown parameter in the above formula.

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \tag{12.1.2}$$

But the value of the substituted standard deviation of the sample is not identical with that of the universe, owing to sampling error; in fact, the former tends to be smaller than the latter because the scatter of values within the sample tends to be smaller than that in the universe. To match this understatement in the numerator (i.e., to maintain the ratio in the fraction), we correspondingly reduce sample size (n) by one in the denominator:

$$s = \sqrt{\frac{\Sigma x^2}{n-1}} \tag{12.1.3}$$

This formula may fittingly be substituted for the population sigma. The correction becomes trivial, of course, if sample size is very large. Nevertheless, it is conventionally included, as a matter of principle, even in large samples. The working formulas, with familiar symbols, will therefore read as follows:

$$s_{\bar{X}} = \frac{\sqrt{\dfrac{\Sigma x^2}{n-1}}}{\sqrt{n}}$$

$$= \sqrt{\frac{\Sigma x^2}{n(n-1)}} \tag{12.1.4}$$

This expression, then, yields the standard error of the mean which we have been looking for and from which we shall construct the interval estimate.

* Greek letters are conventionally used to represent *parameters* (e.g., σ, μ); Latin letters (e.g., s, $\bar{X}$) represent *sample statistics*, which are necessarily estimates. But this practice is not uniformly adhered to by writers in the field.

is even three sigmas removed. But, if we had the value of the standard deviation of the sampling distribution, we could set up an interval estimate for the true mean; hence, we must obtain an estimate of this value. We label this value, *not* a standard deviation (SD), but more descriptively a *standard error* (SE), since the deviations of all sample means from the true mean are actually sampling errors.

Our sample alone must furnish the information on which to base an estimate of the standard error, since it is all we have. Hence, we must explore the sample to uncover some cues to the magnitude of the sought-for standard error. Initially, we might guess that the size of the sample (n) and the degree of scatter in the sample (SD) are related to the size of the standard error, according to the following reasoning.

First, as to sample size, we should expect larger samples to be more representative of the universe than smaller samples, and we would therefore expect larger samples to yield smaller sampling errors. In fact, when a sample contains every last item in a finite universe (a 100 per cent sample), there can be no sampling error at all. Every "sample" mean would then necessarily be identical with the universe mean, and the standard error would be zero. At the other extreme, the smallest possible sample would contain only one case ($n = 1$). The distribution of sample means would then be identical with that of individual universe values, and would display the same amount of variation as the items in the universe itself. Hence, in that instance, the standard error would be equal to the standard deviation of the universe (σ). From this brief analysis, we may conclude that the standard error of the mean (a) will always be smaller than the standard deviation of the universe, since n will always be larger than 1; and (b) will decrease as sample size (n) increases, reaching zero when $n = 100$ per cent of the universe.

Second, we intuitively expect the degree of scatter in the universe, as reflected by the degree of scatter in the sample, to affect the size of the sampling errors. Thus, if all values in the universe were alike, all sample means would also be alike, and there would be no sampling error whatever. The standard error of such a completely homogeneous universe would always be zero! A sample of one boy would give us without sampling error the average number of arms for all boys, since every boy has the same number of arms — a perfectly homogeneous universe; however, a sample of one boy would not give us without error the average height of all boys There is a great variety of heights, which produces comparable variety within samples, and consequently among the sample means as well. Inverting this argument, we conjecture that a high degree of scatter in the sample is indicative of heterogeneity in the universe and therefore of large sampling errors. We thus anticipate that the standard error for a given n will be larger when sampling a very heterogeneous mass of data than when sampling a relatively homogeneous mass.

and from any randomly selected mean we become virtually certain that the resultant interval will trap, as it were, the target mean. Even the most pessimistic soul, who always fears the worst, may give himself a sense of relative security by attaching three standard errors to his observed sample mean.

This method, be it noted, does not disclose the exact whereabouts of *the true* mean, as *has been previously cautioned; it merely furnishes* a stronger or weaker expectation that the true mean will be found within the specified interval estimate, derived from the observed sample. The true mean is stationary, wherever it is. Since it is the interval estimate that either succeeds or fails in enclosing the parameter on a given sampling trial, we may properly speak of the probability of such an interval succeeding in encompassing the universe mean. Of course, once the sample has been drawn and the interval formed — once the trial is over — it either does or does not include the target mean between its calculated limits, although we will never know. Our confidence rating in that interval corresponds to its pre-trial probability, which has been selected for our convenience and purpose.

Working Procedure. To illustrate the process of calculating a confidence interval, we carry out all necessary operations on the following sample of 80 suicide rates:

6.0	18.6	4.6	11.2	18.8
24.9	9.1	14.7	11.6	22.4
13.4	11.5	8.2	28.8	10.0
5.4	7.5	26.1	8.6	8.1
23.4	11.6	14.8	13.6	6.0
8.4	10.0	10.4	9.6	15.3

(1) We compute the sample mean:

$$\begin{aligned} \overline{X} &= \frac{392.6}{30} \\ &= 13.1 \end{aligned}$$

(2) We compute the standard error of the sample mean:

$$\begin{aligned} s_{\overline{X}} &= \sqrt{\frac{1{,}241.2}{(30)(29)}} \\ &= 1.2 \end{aligned}$$

(3) We multiply $s_{\overline{X}}$ by the number of sigma units to be added and subtracted for the agreed-upon confidence interval — for example, 1.96 for the 95 per cent interval:

$$\begin{aligned} 1.96s_{\overline{X}} &= 1.96(1.2) \\ &= 2.4 \end{aligned}$$

Construction of the Confidence Interval (Large Sample). Since the sampling distribution of the mean is approximately normal around the population mean, it follows that two out of three sample means (68.27 per cent) will lie within one standard error of the true mean. Our particular sample mean will therefore have a 2 in 3 chance of falling within one standard error of the true mean. Hence, by adding and subtracting one standard error to and from our randomly selected sample mean, we have an approximately 2 in 3 chance of enclosing the true mean. *Similarly, 95 out of 100* sample means lie within 1.96 standard errors of the true mean; in consequence, *if we attach 1.96 standard errors to either side of the sample mean,* we have a 95 per cent probability of enclosing the true mean. In general, *it is possible to provide any desired confidence interval* by the simple technique of attaching the requisite multiple of the standard error to the *observed sample mean.* *By thus making our interval estimate* wide enough, we may be practically certain that the true mean has been contained within the interval, even though the observed sample mean is highly inaccurate.

Let us suppose that we have drawn a "bad" sample, whose mean, $\bar{X}_o$, is located in the tail of the sampling distribution, far removed from the universe mean (μ) which it intends to represent (Figure 12.1.2). Clearly, in this instance, it would not have been sufficient to attach only one standard error to either side of the observed mean in order to catch the true mean; nor would it have been enough even to attach two standard errors, since the true mean is more than two standard errors distant from our illustrative sample mean. However, if we affix three standard errors to either side of the observed mean, we obtain an interval whose lower limit extends substantially beyond the true mean. *Since practically all* (99.74 per cent) of the sample means lie within three standard errors of the universe mean, by adding and subtracting three standard errors to

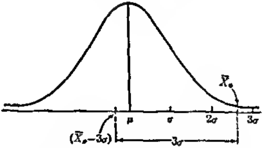


Figure 12.1.2 *Theoretical Sampling Distribution of Mean, Observed Sample Mean, $\bar{X}_o$*

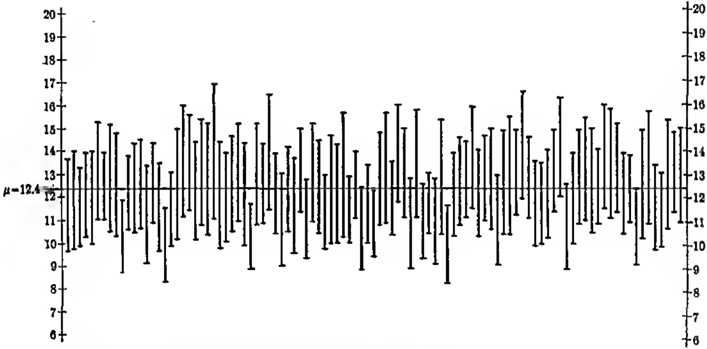


FIGURE 12.1.3 *100 Interval Estimates, 95 Per Cent Confidence, n = 30; Universe: Suicide Rates, 107 U.S. Cities, 1950*

(4) We add and subtract $1.96s_{\bar{X}}$ to and from the observed sample mean, $\bar{X}$. Thus:

$$\bar{X} \pm 1.96s_{\bar{X}}$$
$$13.1 + 2.3 = 15.4$$
$$13.1 - 2.3 = 10.8$$

(5) We finally make the interpretation that the chances are 95 in 100 that the true mean will be found within the interval 10.8–15.4. We are reasonably sure that our interval contains the true mean, since only 5 intervals in 100 constructed in the same manner will fail to do so. If, however, it doesn't, something has happened which would occur only 5 times in 100, a risk which we may be willing to face.

But how decide on the width of the confidence interval? Here one can offer only the most general guidance. Into the choice will enter a consideration of the consequences of the decision, the risk one will wish to run, and even the temperament of the person concerned. In effect, a decision like this is comparable to any other of the numerous decisions we make in the face of life's uncertainties.

Demonstration of Experimental Confidence Intervals. All statements of probability, it will be recalled, are based on the assumption of an infinite number of trials. While it is not possible to conduct such an infinity of trials, it is usually possible to conduct a fairly large number of experimental trials, the outcomes of which then serve to test the initial probability statement. To that end, we have calculated the 95 per cent confidence limits for each of our 100 samples ($n = 30$) of suicide rates, in order to determine whether the resulting intervals would actually enclose the true mean approximately 95 per cent of the times. The diagram on the opposite page (Figure 12.1.3) portrays the locations of the 100 confidence intervals, along with the known population mean of 12.4. It will be observed that exactly 95 out of 100 interval estimates do enclose the true 'mean, a result which is, surprisingly enough, identical with theoretical expectation. Here is tangible evidence that the confidence we place in a properly constructed interval estimate is amply justified.

Convenience of Large Samples. Up to now, all probability statements have been based on the assumption that the sampling distribution of the mean is normal. It is reassuring that such a sampling distribution will always be normal when the universe itself is normal. A normal population produces normality in this sampling distribution, whatever the size of the sample.

However, it is particularly characteristic of sociological data that populations are often not normal. Size of families, income and other social variables distinguish themselves from typical psychological data in that

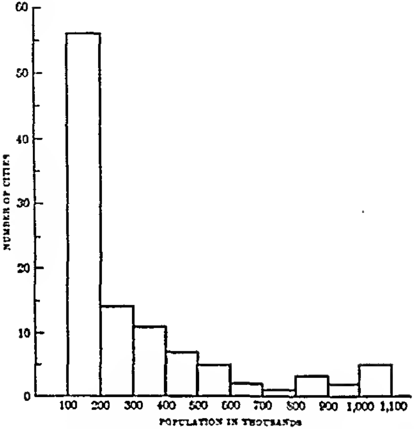


Figure 12.1.4 *Histogram, 106 American Cities by Size of Population*

Table 12.1.Ax

Distribution of 200 Sample Means, n = 20

Mean ('000)	f	Per Cent
100-199	2	1.0%
200-299	50	25.0
300-399	67	33.5
400-499	31	15.5
500-599	18	9.0
600-699	17	8.5
700-799	8	4.0
800-899	4	2.0
900-999	1	.5
1,000 or more	2	1.0
	200	100.0%

they are often severely skewed. We may therefore raise the question of whether such skewness disturbs the required normality of the sampling distribution. The answer is: it does. But at the same time, the effect of such skewness can be circumvented by appropriately increasing the size of the sample beyond the approximate minimum of 30. As sample size increases, the sampling distribution approaches normality. However, no general statement can be made about the rate at which the sampling distribution approaches normality, since this rate will vary according to the severity of the skew to be offset. But we may gain some insight into what might on its face seem a rather remarkable phenomenon by examining the following experiment, in which we drew samples of size 20 and samples of size 40 from a compilation of 106 large American cities for size of population. The original distribution (Table 12.1.3) was highly skewed and

Table **12.1.3**

Population Distribution, 106 American Cities 100,000 and Over, by Size of Population

SIZE OF CITIES ('000)	*f*	PER CENT
100–199	56	52.9%
200–299	14	13.2
300–399	11	10.4
400–499	7	6.6
500–599	5	4.7
600–699	2	1.9
700–799	1	.9
800–899	3	2.8
900–999	2	1.9
1,000 or more	5	4.7
	106	100.0%

Source: U.S. Bureau of the Census, *U.S. Census of the Population: 1950, Characteristics of the Population*, Vol. II, Part 1, U.S. Summary, U.S. Government Printing Office, Washington, D.C., 1953, Table 1b.

consequently ideally suited for experimentally testing the effect of a skewed population on the sampling distribution of the mean.

The first set of 200 samples ($n = 20$) was drawn and the mean of each sample calculated and tabulated (Table 12.1.4a). This experimental sampling distribution is still perceptibly skewed to the right, but not nearly to the same degree as was the parent universe. This reduction in skewness is a result of the fundamental principle that the means are great levelers which necessarily cut down the individual extremes found in the universe. It follows therefore from the very nature of the mean that a distribution of sample means can never be as irregular, or as widely dispersed, as are the individual values which make up the

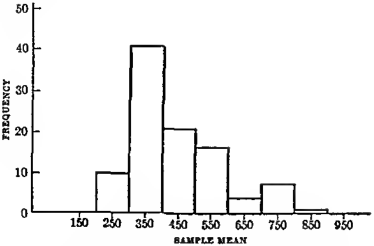


FIGURE 12.1.5h *Histogram, Distribution of 100 Sample Means, n = 40*

applying normal probability calculations to their data by taking proper precautions.

The convenience of large samples consists in the assurance that the sampling distribution of the mean is approximately normal, an assumption which may be in doubt when samples are small. Situations do, however, often arise when small samples are not only convenient, but even unavoidable. In such instances, modification of the foregoing procedures is required, although no new principles are involved. This adaptation of reliability techniques to small samples is therefore not considered in this text.

Confidence Interval for a Percentage (Large Sample). Like the mean, the sample percentage also raises the issue of reliability, since sample percentages are equally subject to sampling errors. Thus, a given sample percentage (e.g., 60 per cent favoring military training) may be off a trifle, or may deviate considerably from the true percentage. Similarly, 30 per cent of a random selection of married women may bè gainfully employed, but the percentage in the sampled universe may conceivably be as low as 20 or as high as 40. Fifty-one per cent of the sample electorate may signify its intention of voting for the Republican nominee in the forthcoming election, but unless we can in some way assess the accuracy of that estimate we cannot forecast with any degree of confidence the outcome of the election. Hence, in interpreting a sample percentage, we must again estimate the average sampling error (standard error) in order to provide a confidence interval.

samples. Furthermore, the larger the sample, the more successfully will the dispersion be reduced. Thus, when sampling from a *markedly* skewed population, samples even as small as 20 will produce a sampling distribution displaying evidence of a strain toward normality, although obviously *falling far short of that goal*

Samples of 40 should therefore continue the trend toward normality, or at least toward symmetry. Such a distribution is tabulated in Table *12.1.4b. Gone now is the long tail; instead, the beginnings of symmetry* are clearly visible. Thus, sociologists who are often called upon to deal with skewed *populations may draw considerable comfort from the* important implications of this little experiment and may feel justified in

Table 12.1.4b

Distribution of 100 Sample Means, n = 40

Mean ('000)	f	Per Cent
100–199	0	0.0%
200–299	10	10.0
300–399	41	41.0
400–499	21	21.0
500–599	17	17.0
600–699	3	3.0
700–799	7	7.0
800–899	1	1.0
900–999	0	0.0
1,000 or more	0	0.0
	100	100.0%

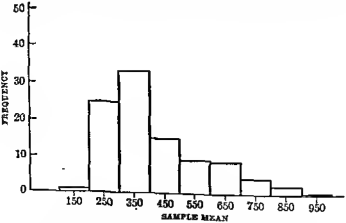


Figure 12.1.5a *Histogram, Distribution of 200 Sample Means, n = 20*

Since intervals constructed in this manner contain the true percentage 95 times out of 100, we may declare that we are confident at the 95 per cent level that the *true percentage is not less than* 56.81, *nor greater* than 63.19. Since 5 per cent may be assumed to exclude the true percentage, we thus take a 5 per cent risk of being wrong.

The foregoing procedure is another instance of large-sample technique in that it applies to samples having 30 or more cases. But the assumption of a normal sampling distribution hinges on the ratio, P:Q, as well as on sample size. As the P:Q balance departs more and more from 50:50, the sampling distribution becomes increasingly skewed.

As has already been pointed out in the study of the mean, in the case of percentages, the skew in the P:Q balance communicates itself to the sampling distribution. Again, as with the mean, this skew can be circumvented only by a sharp increase in the size of n, so as to instill normality in the sampling distribution. Table 12.1.5 shows the sample sizes needed for varying sample percentages in order to permit the assumption of a normal sampling distribution.

Table 12.1.5

Minimum n for Selected p-Values for Normal Approximation

p	n
.5	30
.4	50
.3	80
.2	200
.1	600
.05	1,400

Adapted with permission from William G. Cochran, *Sampling Techniques*, p. 41. Copyright 1953. *John Wiley & Sons, Inc.*

The Finite Population Multiplier. In actual sociological investigations, we ordinarily encounter finite populations, and yet in the foregoing instances we have calculated the standard error on the assumption of sampling from an infinite universe. But this is not quite so unrealistic as might appear on its face. Finite populations are usually large enough to be considered as infinite, which is incidentally something of a practical convenience, since reliability procedures are somewhat less complex for infinite than for finite populations. For finite populations, the measurement of reliability must take into account both sample size (n) and the size of the universe (N), whereas for infinite populations, the computation of reliability need take into account only sample size, since the size of the universe is incalculable and hence cannot vary.

The technique of constructing an interval estimate of a universe percentage is in its essentials no different from that of the mean: we (1) fix the desired degree of confidence; (2) *estimate the standard* error; and (3) attach to the observed sample percentage the multiple of the standard error needed to attain the agreed-upon confidence. The sampling theory on which this procedure rests is identical with that underlying the treatment of large sample means: namely, the distribution of *all possible percentages (p) is approximately normal* around the universe percentage, P, with standard error

$$\sigma_p = \sqrt{\frac{PQ}{n}} \tag{12.1.5}$$

where σ_p = standard error of percentage
P = universe percentage
$Q = 100 - P$

But since the parameters P and Q are of course unknown, we again follow the convention of substituting the sample values p and q in order to obtain an estimate of the standard error: *

$$s_p = \sqrt{\frac{pq}{n}} \tag{12.1.6}$$

Computing an Interval Estimate of a Percentage. Let us suppose that out of 900 randomly selected high school students, 60 per cent respond in favor of universal military training and 40 per cent are unfavorable. To form a confidence interval, we must first estimate the standard error:

$$s_p = \sqrt{\frac{(60)(40)}{900}}$$
$$= \sqrt{2.667}$$
$$= 1.63$$

By adding and subtracting this quantity to and from the sample percentage, we obtain the 68 per cent confidence interval:

$$60\% \pm 1.63 \text{ percentage points}$$
$$58.37\% - 61.63\%$$

The odds are 2 to 1 that the true percentage lies within this interval. If we should desire greater confidence — say, 95 per cent — we would attach 1.96 standard errors:

$$60 \pm 1.96(1.63)$$
$$60 \pm 3.19$$
$$56.81\% - 63.19\%$$

* As in the case of the mean, a better estimate of the standard error of p is provided by $\sqrt{\frac{pq}{n-1}}$. For a discussion of this point, see William G. Cochran, *Sampling Techniques*, John Wiley & Sons, Inc., New York, 1953, pp. 32–33.

gardless of the size of the universe, that has given birtb to the paradox that sample reliability is affected only by the absolute number of cases and that the sampling ratin is of nn eonsequence.

Since tbe sampling ratin is often relatively small, as in public opinion surveys, the finite population multiplier is correspondingly ignored, which explains its textual omissinn in some nf tbe briefer discussions of sampling. Howevcr, it should be noted that, wbenever the universe is extremely small, as a small community, tbe sampling ratio must necessarily be large, otherwise tbe number of cases in tbe sample would not be adequate to represent tbe sampled universe. Thus, tbe U.S. Bureau of tbe Census has seen fit to draw a 25 per cent national sample in order to ensure adequate coverage in the very small enumeration areas. A mucb lower sampling ratio would nf eourse be sufficient for national, state, and metropolitan coverage, but it would produce an insufficient number of sampling units from the numerous tiny subdistricts.

Some Persisting Issues in Applied Sampling

The complete process of estimation as exemplified in the preceding discussion is easily reducible to a few abstract stages: (1) the drawing of tbe sample, (2) the computation of the sample statistic, and (3) the formulation of the inference concerning tbe parameter.

But sampling is not an abstraction; it is not carried out in a vacuum. It is performed on concrete things and events: for example, a campus of students, a carload of corn, a community of households, a set of opinions on public issues, or almost any conceivable set of events about which there is some scientific curiosity. Being an operation executed in material situations, its validity is conditioned hy the nature of the data it works on. What are the characteristics of the social data which tbe sampler would have to take cognizance of, in order to insure sound statistical reasoning?

The Instability of the Social Universe. The stages of the sampling procedure sketched above seem quite uncomplicated wben applied to a bolt of cloth, batches of cbemicals, a shipment of light hulbs, or the leaves of a mimosa tree. The inferences are made, in general, as quickly as the data can be processed, and the universe is not subject to disturbing accretions, deletions, or deterioration. However, such uneventful simplicity is not descriptive of the social world, which is mucb more complex and dynamic. Social investigations are frequently of long duration; the universe, the elements of which are very mobile, is often subject to alteration by growth as well as diminution, and its significant traits are subject to modification while nnder stndy — indeed, partly as a result of their being studied. Thus, pnhlic opinion is subject to vacillation, This principle is in agreement with common sense: a sizable fraction of a finite universe should he more representative than a negligible fraction thereof. A 100 per cent sample would be most dependable, since in such cases there can be no sampling error at all. The standard error would he zero. Analogously, we expect a 50 per cent sample to be more reliable than a 25 per cent sample, which in turn we expect to yield more confidence than a 5 per cent sample. In general, relatively large samples are more representative than small ones. The measurement of this improvement in reliability is effected hy incorporating the *sampling ratio, n/N*, in the formula for the standard error. Thus, the standard error of the mean, on the assumption of a finite population is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} \qquad (12.1.7)$$

The term, $1 - n/N$, or the proportion of the population not in the sample, is frequently referrsd to as the *finite population multiplier*,* because it measures the improvement in reliability attributahle to the finite nature of the universe.

It is clear that the quantity, $1 - n/N$, will always be less than 1.00; hence, it will always produce a shrinkage in the standard error below that for the infinite population. When the sampling ratio is large, it will even lead to a substantial reduction in the size of the standard error. Thus, with a 25 per cent sample, such as that selected by the U.S. Bureau of ths Census in 1960, employment of this term would lead to a reduction in the standard error as follows:

$$\begin{aligned} fpm &= \sqrt{1 - .25} \\ &= \sqrt{.75} \\ &= .86 \end{aligned}$$

Eighty-six per cent of the initial *SE* constitutes a 14 per cent shrinkage. Hence, it would here be imperative to employ the multiplier, otherwise we would fail to do justice to the accuracy that actually resided in the sample data. On the other hand, when the sampling ratio is small, say less than 5 per cent, the finite multiplier has little or no effect on the magnitude of the standard error and may be disregarded. Thus, in a sample of 100–200 cases it makes little difference whether the universe contains 10,000, 1,000,000, or even 180,000,000 units. It is this fact that several hundred sample observations may be equally reliable, re-

* More commonly referred to as the *finite population corrector*. However, this term is somewhat misleading. The factor, $1 - n/N$, does not serve to "correct" a faulty result; rather it is an integral component in the standard error formula for finite populations. When the sampling ratio is small, this component may be sloughed off, whereupon the standard errors for finite and infinite populations become identical. If the standard error had been given originally for finite populations, then it would have been necessary to "correct" for the infinite population!

example — the further use of the sample touches a population which did not have "an equal chance of being selected." Such prolonged use of a sample beyond its normal life is analogous to the retention of outmoded clothing or dilapidated household utensils — they may still serve their purposes, but not too well. Anyway, we cannot discard them because we cannot afford new ones. In view of this dilemma, we are obliged to coin new terms and supplementary concepts which do not cohere exactly with the ideal simple random procedure: the "target universe," "hypothetical universe," and the "availability sample" which we project to the universe which did not originally have an opportunity to contribute to the sample.

The lesson we can derive from contemplating the conditions just described is twofold: (1) the interpretive inferences in social statistics necessarily take on a less rigorous character and must be made with caution and reserve, and (2) the requirement of familiarity with the subject matter is all the more insistent, without which the necessary sophistication and skill cannot be brought to bear on the application of quantitative reasoning.

Size of Sample. The inevitable question that arises early in any piece of investigation is that of the size of sample. The taking of any sample, in preference to the entire universe, is of course an economy. But there is no economy in a sample that is larger than necessary; and it is pound-foolish to be content with a sample that is too small to yield the requisite accuracy. The importance of this problem of sample size is much greater than one would deduce from the meager discussions available in elementary texts. And yet the question is bound to arise in even the simplest sampling studies. The techniques appropriate to answering it are matters of fairly advanced statistics. However, some broad principles are at hand which will serve to pilot the student who is not equipped with more delicate instruments.

To the novice, it usually seems incredible how small a sample may often produce dependable results. Literally, a few thousand interviews have successfully predicted the outcome of U.S. national elections. How is that observation to be reconciled with the foreboding description of the unstable social world?

The most immediate explanation of this apparent incongruity is that even the social world may be viewed at times in simple dimensions. A political poll, in a system comprising only two parties, and in which the voters are sampled according to well-established criteria, might very well yield highly reliable results — when there is no unforeseen intervention of novel circumstances. As a touchstone to our statistical judgment of sample size, we should visualize the supremely uncomplicated situation of a perfectly homogeneous universe which, of course, could be sampled and some of the samplings of the decennial census are touched with obsolescence almost before they are circulated for use. The berpectacled scholar stands contemplating his sample while the social universe which it supposedly represents is moving on. Pre-election polls are therefore taken at short intervals to eliminate the pitfalls of opinion shifts; but U.S. decennial censuses are too cumbersome and costly to replicate at useful intervals. Hence, we sometimes find ourselves helplessly falling back on an anachronistic population base in our sociological studies.

This dilemma is aggravated by the fact that, in the social studies, the unit of time over which the social process endures is usually much longer than it is in the studies of the physical world. A meteorologist, who would not predict the weather more than a few hours in advance, expects an economist to forecast depression and the sociologist to predict the course of a delinquent career and to foresee recidivism.

The Heterogeneity of the Social Universe. The instability of the social world, as previously described, constitutes heterogeneity in the chronological dimension: divorces, crimes, and marriages are not at all identical with those events, which are nominally the same, of a decade previously. Public opinion today is not what it was "before the last diplomatic note from Russia."

But heterogeneity also exists in the lateral dimension. The superorganic world, with its variety of personality and cultural traits, does not offer such a homogeneous base for reliable sampling. Subcultural areas within the metropolitan community display such culturel diversity as to render sample designs complex and difficult to execute. The Elmtowns, Yankee Cities, and Middletowns, with all their presumed typicality, cannot be cavalierly employed as launching pads for inferences about other Elmtowns, Yankee Cities, and Middletowns, even though the economy of sociological sampling enterprises sometimes compels us to do so.

Sociologists might well envy the students of the inanimate world or even the botanical and animal kingdom, who preside over a relatively homogeneous universe: a carload of wheat or light bulbs, a sample of dogs or mice in the laboratory But such is not the fortune of the social scientists. Hence, as an economy measure, we are often obliged to employ obsolete social samples that should have been cast aside for more current samplings. They are still more often applied to "target" populations, which were not even included as components in the original universe from which the sample was selected.

As an obvious consequence of the practical considerations recited above, the ideal model of random sampling often becomes impossible to carry through. Even if it is practiced in a given study — in the census, for sample is impracticablé, as in national opinion polling, stratification may perhaps be utilized in order tn reduce it.

Work to Be Performed by the Sample. The objective of any sample is to represent the unknown universe with a certain desired precision. But it can perform this task either well or badly. The principal statistical criterion of a well-behaved sample is, nf course, the size of the standard error: all other things hcing equal, a larger sample will produce a smaller standard error, a smaller sample will produce a larger average sampling error. It should be unnecessary tn reiterate that this principle will generally hold only with random samples; for non-random hiased samples, there is no safety in numbers, as the calamitous experience of the *Literary Digest* poll of 1936 has amply testified. Like a fleet athlete running in the wrong direction and reaching the wroog goal so much sooner, an increase in the size of a hiased sample can lead only to an all the more grievous hlunder.

In the second place, if the sample items are to he suhclassified into smaller categories, the sample obviously has to be large enough to ahsorb such dismemberment without incurring the hazards of vacant or thinly populated cells. If some of these categories show sparse frequencies, reflecting a very unhalanced trait distrihution in the universe (e.g., delinquents, left-handed persons, Negro physicians), the over-all sample will necessarily have to he adequate to catch such infrequent items.

Furthermore, the prohlem of size of sample would he much less formidahle if the task of a sample were confined to one variahle. But this is rarely the case. We know that a sample can he a very costly instrument. Like any other investment, it is economical to make it do the work for as many purposes as possihle. Thus, a given sample of several hundred household units may be used to survey socio-economic status, religion, size of family, unemployment, or any other variable within the scope of the surveyor's scientific curiosity. Not only may samples he employed in this multi-purpose sense, but a sample usually contains many subsidiary characteristics which may be included as hackground information in the schedule. The importance of this diversity of work tasks is that a sample of a given size will not be equally efficient for all variables under study. For the simple dichotomy of sex, the sample may he representative enough, whereas for school attendance, unemployment, and many other more complex distributions, the same sized sample will be much less representative. Even the most elementary surveys will be called upon to take into consideration the wide differences in patterns of distribution.

Because traits are not equally variahle, a sample that satisfies the demands of a two-value trait will probably not be copious enough for a more complex distrihution. If such multi-dimensional tasks are planned by a single case. Such a sample cannot be spoiled by even the most arbitrary selection. This is the reason why many an otherwise careless and inadequate poll does sometimes succeed in doing its work well and is bound to make even an incompetent surveyor look good.

On the other extreme, a hypothetical aggregate of 1,000 totally unique objects could not be successfully sampled at all. Not even the most scrupulous randomization, nor any enlargement of sample size, would ameliorate this impossible circumstance. In practice, of course, neither one of these abstractly perfect extremes is ever encountered. Our populations always fall somewhere on the long continuum between these theoretical limits.

Consequently, the question cannot be so simply put as: "What size sample?" In such an oversimplified form, it can be answered only evasively: "It all depends." On what does it depend? It depends on three interlocking circumstances: (1) the presumable characteristics of the universe which the sample is designed to represent faithfully; (2) the work required of the sample in performing its function; and (3) the material resources at the disposal of the investigators.

This last named consideration is, of course, extraneous to the question of statistical principles. Nevertheless, the cost per case is an important budgetary issue in all research and will be a final determinant in the compromise between the optimum sample that we would like and the practical sample that we can afford. Beyond this acknowledgment of the problem of cost accounting in sampling, this issue will not be further discussed.

Characteristics of the Universe. Although, in theory, the characteristics of the universe are unknown, in actuality a seasoned worker knows quite a lot about the universe that he is about to study for a specific trait. Since it is an exaggeration to state that the universe or a parameter is *literally "unknown,"* the general characteristics are indeed taken into account by the statistical worker.

Thus, the more heterogeneous the universe is judged to be, the larger the sample should be. If there were no variation, there would be no necessity to sample more than one item. By corollary, a wide range of variation in quantitative data requires a correspondingly larger sample to assure representativeness. The same principle would obtain in the case of qualitative variables. As has already been demonstrated (Chapter 7), an even division in qualitative variables yields maximum heterogeneity, which is measurable by the standard error of a proportion: an 80-20 split would give a smaller standard error than a 50-50 division. Thus, to assure a given reliability on a Yes-No vote, an expectation of an approximately even split of votes would demand a larger sample than would an 80-20 division. If, for any reason, a large simple random

4. (a) For a given population, determine how sample size would have to be adjusted in order to reduce the standard error of the mean to:
   one-half its original size
   one-fourth its original size
   one-sixteenth its original size
   (*Hint:* Substitute in *SE* formula and solve.)

   (b) Explain how these results illustrate the statement that the reliability of the mean varies directly with sample size.

5. A given sample has the following characteristics:

$$\bar{X} = 11$$
$$s = 3$$
$$n = 100$$

If the population mean is known to be 12, what is the sampling error of the observed mean? What is the estimated standard error of the mean? How often would you expect sampling errors larger than ±.5? ±1.00? (*Hint:* Find the required *z*-measure and refer to the table of normal areas.)

6. Explain the statement that the sampling distribution of the mean is the equivalent of the distribution of sampling errors.

7. Explain in your own words why the sampling distribution of the percentage will be lacking in symmetry when n is small and $P{:}Q$ very unbalanced — say, 90:10. (*Suggestion:* Draw an appropriate graph with the percentage scale on the base line.)

8. Suppose that a complete set of attitude scores has a mean of 1,000 and a standard deviation of 200.

   (a) If 25 scores are picked at random, what is the probability that their average score will be less than 950?

   (b) Greater than 1,100?

   (c) Between 900 and 1,100?

   (*Hint:* Obtain the *SE* and *z*, and consult table of normal areas.)

9. Discuss the statement: "The wider the interval estimate, the greater the degree of reliability; the greater the degree of precision, the less reliable the interval estimate."

10. (a) State in what sense the standard error is an average of all possible sampling errors.

    (b) Will an analogous statement hold for the standard deviation?

11. Assume a sample survey of a college student body on the following characteristics: (1) proportion who read the newspapers regularly; (2) attitude toward segregation; (3) opinion on rules for women's residence halls; (4) proportion who receive A grades; (5) proportion who attend church regularly; (6) career objretives; (7) social class of parents; (8) proportion gainfully employed and their earnings. Would a single sample serve to represent all of the above characteristics equally well? Why or why not?

for a given sample, the sample has to be so designed to meet the peak requirement of the most demanding individual piece of investigation. Such an all-purpose sample, for instance, is supplied by the U.S. Census. The sample size of 25 per cent, set for the 1950 enumeration, is large enough to satisfy any sampling need that can be anticipated — from the smaller town to the giant metropolis.

The statistical sample thus constitutes a major road to quantified knowledge. If it never gives us certainty, but only probabilities, it shares this probabilistic quality with knowledge of every type. It simply is not given to man to be completely ignorant, any more than he is omniscient or omnipotent. This is another way of saying that statistics is a truly human science — a tool in the pursuit of knowledge that will never terminate.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Statistical Inference
   Estimation
   Parameter
   Statistic
   Sampling Error
   Sampling Distribution
   Confidence Interval
   Standard Error of the Mean
   Standard Error of the Percentage
   Interval Estimate
   Point Estimate
   Large Sample
   Finite Population Multiplier

2. *Using the table of random digits (Table IV, Appendix), draw two samples of* 30 cases each from the list of 107 suicide rates (Table 3.1.1a).
   (a) Compute the mean and the standard deviation of each sample.
   (b) Compute the sampling error of each sample mean on the basis of the known true mean.
   (c) As a class project, arrange the means from (a) above in a frequency table. Compare this distribution with the empirical distribution presented in the text. Comment on the difference.

3. (a) Estimate the standard error of the mean from each sample of 30 cases from the preceding exercise.
   (b) Explain why the estimated standard errors are not identical in value.
   (c) Establish 95 per cent confidence intervals by adding to and subtracting from each corresponding sample mean 1.96 standard errors.
   (d) As a class project, determine what percentage of the 95 per cent confidence intervals contain the true mean.

then, does the null hypothesis differ from any other hypothesis — from a hypothesis, for example, stating that juvenile delinquency is associated with status discontent, or that the incidence of suicide depends on the degree of social disorganization? These hypotheses we may call explanatory hypotheses, whereas the null hypothesis is an auxiliary device which tentatively denies the validity of the explanatory theory. Hence, if we reject the null hypothesis, we strengthen the credibility of our explanatory hypothesis. But the null hypothesis itself has no explanatory value.

Yet science is in search of relationships. Hence, in accordance with the above reasoning, the null hypothesis is usually launched with the expectation, indeed with the hope, that it will be nullified, as the derivation of the term implies. Nevertheless, it represents a possibility that must be disposed of before alternative hypotheses which imply assignable causes can be considered.

Returning to the comparison of eastern and western cities, we would set up such a comparison in the first place only in order to uncover some of the factors that determine the incidence of suicide. If we do discover a significant difference between the average rates of East and West and consequently reject the null hypothesis, we will have added a small increment to our knowledge of suicide — namely, the regional factor as a source of variation in the suicide rate. If, on the other hand, the two samples of data show no significant difference and we accept the null hypothesis, then we have not advanced our understanding, although supposedly we will not have retrogressed. In this latter case, we would have to contrive new kinds of comparisons in our search for the assignable causes of suicide.

Thus, it is only when the preliminary hypothesis of "no difference" has been cleared away that we gain some insight into the occurrence of suicide. Hence, the null hypothesis is inherently linked with a more constructive statistical hypothesis, sometimes called an alternative hypothesis, which becomes tenable to the extent that the null hypothesis has been discredited. Such an alternative hypothesis may specify an exact degree of difference between East and West, with a view to gauging the strength of the unknown variables which produce suicide in a disorganized society. Clearly, there is no point in pursuing the effects of a given control or experimental variable such as geographical region if we have found in favor of the null hypothesis. Thus, by analogy, the null hypothesis serves the purpose of a criminal trial: we set up the hypothesis of innocence, giving the evidence an opportunity to nullify it. Only if and when that presumption is rejected does the court give thought to alternative punishments corresponding to the degree of guilt.

In accordance with the foregoing principle, the null hypothesis has come to be predominantly identified with two types of research pro-

SECTION TWO

Testing a Statistical Hypothesis

Hypothesis-Testing and Estimation Compared. There are two general types of statistical inference: **(1)** ***estimation***, which begins without any stated assumption about the value of the parameter and merely seeks to estimate descriptively what the parameter could be; and (2) ***hypothesis-testing***, which begins with a hypothesis about the parameter and then uses the sample data to check the plausibility of that statement.

In the previous section, we were concerned with problems of estimation. We began, for example, with the observation of a sample mean, and from this we derived an interval estimate with a specified degree of confidence. We first drew our sample and then made our estimate of the parameter.

But in hypothesis-testing, we formulate our hypothesis about a parameter in advance of the collection of the sample data, which is then used to test that hypothesis. We may, for example, hypothesize that the average suicide rate of eastern cities does not differ from that of western cities. We begin with that supposition, and then we take an appropriate sample of eastern and western cities, compare the two means in the prescribed manner, and finally reach a decision whether, in the light of the sample difference, the hypothesis should be accepted or rejected.

Although an interval estimate is always derived from the previously compiled sample data, a statistical hypothesis involves quite another analytical process. This process starts with an antecedent conjecture about an unknown population value, presumably arrived at without benefit of the undrawn sample. Furthermore, a hypothesis is tested and then acted upon: it is either accepted or rejected. It requires a decision, which is made with a certain degree of confidence, and which is either correct or incorrect after it has been made. The emphasis thus shifts from mere estimation to ***decision-making***.

The Null Hypothesis (*H_0*). It has become a convention in statistical testing to open the investigation with the null hypothesis, sympolized H_0 In its most current usage, this hypothesis holds that two or more given samples have come from statistically identical populations and that therefore any observed difference between such samples is merely a chance variation. The aforesaid hypothesis that East and West do not differ in their average suicide rates would therefore be a typical null hypothesis.

Essentially, the null hypothesis is set up to be nullified; however, every other type of hypothesis is also set up for that possibility. How, ative of an actual difference between the means of the sampled populations, or whether it could more plausibly be accepted as mere sampling variation. By rule of thumb, we lay down the null hypothesis that the population means of East and West are identical and subject this hypothesis to the statistical test. If we reject this hypothesis, we then proceed as thougb there were a real difference, whereas acceptance presumably implies that the two populations are alike and that the observed difference is a mere sampling error.

An important principle of our testing procedure is that *we can never prove the null hypothesis true.* The best possible evidence that we could ever obtain for the identity of the two population means would be an identity between the two sample means. But even if samples of East and West were to show identical averages, we still could not be positive that the null hypothesis was true. Such an identity between sample means could itsell very well be the result of sampling errors, because population means could differ by a sizable amount and the sample difference still be zero. Similarly, the finding that a sample of married spouses are of the same average age would not prove that all married couples show equal averages. From sample observations, the conclusions we draw about the truth of the null hypothesis are necessarily of a probabilistic nature.

If we can never prove the null hypothesis true, may we prove it definitely false? Here the statistical evidence may be more convincing. The best possible evidence that the two populations differ would be a difference between samples. A small difference would not be very compelling; however, as this difference becomes larger and larger, the case for a population difference becomes stronger and stronger. If the difference hetween samples becomes so large that it is highly improbable that they stem from the same universe, then we may with practical assurance reject tbe null bypothesis. Even here, however, there is no infallible certainty, since the one case in a million — the freak event — may happen. Highly improbable events are regularly occurring, however startled we are whea they befall us. But such an extreme case — the one in a million — is so improbable that most of us would discount it hy regarding it as impossible, thereby rejecting the null hypothesis with great confidence. In other words, we are willing to reject the null hypothesis when the statistical probabilities of being wrong are small enough to suit our purposes.

These decisions — whether or not to reject the null hypothesis — are accordingly made with varying degrees of confidence. This confidence varies accordiug to the probability that a difference at least as large as that observed could have beea obtained by chance, assuming the truth of the null hypothesis. We must therefore now turn to the procedure hy which such a probability is determined.

cedures: (1) *the comparison of two or more populations* on a given trait and (2) the correlation between two or more traits in a given population. In the first type, the null hypothesis posits no difference between population parameters; in the second type, it asserts a chance relation, or zero correlation, between the variables under study.

Principles of Testing the Null Hypothesis. Let us now suppose a random sample of 30 eastern cities and a comparable sample of 30 western cities whose mean suicide rates are 10.4 and 14.3, respectively (Table 12.2.1). We wish to know whether the observed difference of 3.9 could be indic-

Table 12.2.1

Suicide Rates, Eastern and Western Cities, $n = 30$

East	West
9	26
17	12
12	12
12	29
16	12
7	12
11	6
8	10
19	19
10	15
10	26
5	11
10	6
8	11
6	9
11	14
9	15
10	15
10	20
7	10
12	15
14	10
9	27
8	8
15	20
5	9
5	12
15	12
10	15
12	12
$\Sigma = 312$	$\Sigma = 430$
$\bar{X} = 10.4$	$\bar{X} = 14.3$

value is sufficiently extreme, or improbable, to justify the rejection of the null hypothesis. To put it quantitatively: how many times out of 100 could we expect to obtain this, or a larger, sample value if the actual difference is zero?

To obtain this *Pr*-value, we simply convert the observed difference into a standard deviate, or *z*-measure, and consult the table of normal areas. Since the mean of the sampling distribution is by null hypothesis zero, we need only divide the estimated standard error into the observed difference to obtain *z*, which tells us how many sigma units our observed difference lies from the hypothetical mean of zero. In symbols:

$$\begin{aligned} z &= \frac{(\bar{X}_1 - \bar{X}_2) - 0}{s_D} \\ &= \frac{D}{s_D} \qquad (12.2.3) \end{aligned}$$

This *z*-measure is generally called the *significance*, or *critical, ratio*, since its magnitude determines whether we judge the observed difference to be statistically significant — that is, whether the probability of the critical ratio is small enough to reject H_0.

To illustrate the calculation of such a critical ratio, we process the data of Table 12.2.1, in accordance with the requirements of Formula 12.2.3.

	East	*West*
$\bar{X}$	$\frac{312}{30} = 10.4$	$\frac{430}{30} = 14.3$
s^2	$\frac{369.2}{29} = 12.7$	$\frac{1{,}058.7}{29} = 37.5$

(1) We compute the difference between sample means, attaching no sign, since our interest lies in the absolute magnitude of the difference irrespective of direction.

$$\begin{aligned} D &= 10.4 - 14.3 \\ &= 3.9 \end{aligned}$$

(2) Compute the standard error of the difference:

$$\begin{aligned} s_D &= \sqrt{\frac{12.7}{30} + \frac{37.5}{30}} \\ &= \sqrt{1.67} \\ &= 1.3 \end{aligned}$$

The Sampling Distribution of the Difference Between Means. In the chosen example, this procedure rests on the sampling distribution of the difference between two sample means, which here functions as did the sampling distribution of the mean in estimation. To define and illustrate this concept, we may imagine a sampling experiment in which we alternately sample from two well-defined universes, selecting a large random sample of size n_1 from one population, and then a large sample of size n_2 from the other. By replacing samples, we may continue to draw samples in this manner indefinitely. For each pair of independently selected samples we compute the difference between the two sample means, amassing in this way as many differences as there are pairs of samples. We now pose the problem: how will these hypothetical differences between paired means distribute themselves, and what will their mean and standard error be? According to sampling theory, an infinite supply of such differences — labeled the *sampling distribution of the difference* — will have an approximately normal distribution, with its mean equal to the true difference between the universe means, and standard error expressed as follows:

$$\begin{aligned}\sigma_D &= \sigma_{\bar{X}_1 - \bar{X}_2} \\ &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (12.2.1)\end{aligned}$$

The estimated standard error will then be:

$$s_D = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (12.2.2)$$

Thus, the standard error of the difference between sample means, like the standard error of the mean itself, reflects both the degree or variation within the sampled populations and the respective sample sizes. The smaller the population variances, and the larger the sample n's, the smaller will be the standard error of the difference.

Testing Procedure. In testing the null hypothesis, the observed difference is treated as a single value in a normal sampling distribution whose mean is zero (indicating no difference between the population means), and whose standard error is estimated according to Formula 12.2.2, above.

We should recall that the sampling distribution gives us the range of all possible sample differences. Therefore, all sampling values within *that theoretical* range are *possible*, but the extreme values in the tail of that distribution are, of course, much less *probable*. Within this range, we now wish to find the probable location of the single difference which we have observed, so that we may decide whether this sample a purely random manner. Still more confidently would we reject the null hypothesis of no difference between East and West, since the P_r-value is only .0027, and tacitly accept an alternative of some difference.

In making his decision, the worker could set up any level of significance which he thought prudent as the threshold at which to reject the null hypothesis. This means that he could set up any degree of risk which he is willing to assume: .001, .01, .05, .10, or even .50, according to the circumstances. But the agony of making a fresh selection of the level of significance in each instance would be a painful business. Hence, the statistical worker finds welcome relief in the .05 convention, which prescribes that the null hypothesis is automatically to be rejected whenever the probability of being wrong in that decision is 5 per cent or less.

We may graphically illustrate the foregoing argument by the theoretical sampling curve, which has, under the assumption of the null hypothesis, a mean of zero. The shaded rejection zones are located in the tails of the normal curve, and are bounded by ordinates erected at the chosen significance level. By a decision arrived at prior to the sampling itself, the null hypothesis is rejected whenever the sample observation falls in the rejection zone, otherwise it is accepted.

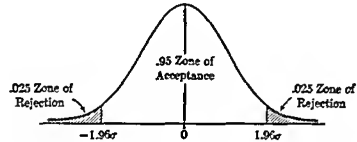


FIGURE 12.2.1 *Sampling Distribution of Difference, $\mu_1 - \mu_2 = 0$, .95 Acceptance Zone and .025 Rejection Zones*

Now, whenever we reject the null hypothesis, we tacitly accept some unspecified alternative hypothesis which has its own sampling distribution. Although any observed sample value is necessarily consistent with the null hypothesis, it is also consistent with innumerable alternatives, any one of which when tested may be more acceptable than the null hypothesis. The truth of this assertion is readily demonstrated by Figure 12.2.2. The farther our observed difference is from the zero mean of the null distribution, the nearer it necessarily is to some alternative which is more likely to be true. We may therefore pose the dilemma:

(3) Express the observed difference as a standard measure, which is the sigma distance between our observed difference and the assumed mean of zero:

$$\begin{aligned} z &= \frac{D}{s_D} \\ &= \frac{3.9}{1.3} \\ &= 3.0 \end{aligned}$$

(4) Entering the table of normal probabilities (Table I, Appendix) with this *z*-measure, we find that .9973 (99.73 per cent) of the values in a normal frequency distribution lie between the mean and ±3 sigmas. Therefore, only .0027 (27 out of 10,000) of all possible sample differences will be larger than the one obtained, assuming the truth of the null hypothesis.

Decision-Making. We must now decide on the basis of this *Pr*-value (.0027) whether or not to reject the null hypothesis. If, on the incomplete evidence of the sample, we too glibly accept the assumption of no difference between the universe values, we run the risk of overlooking a genuine difference. If, on the other hand, we reject the assumption of H_0 too hastily, we run the risk of inferring a significant difference where none may actually exist. How, then, do we decide?

There are two related steps in the process of statistical decision-making which must be differentiated: (1) an evaluation of the obtained probability, and (2) an assessment of the consequences of a wrong decision. In the above calculation, the observed difference between East and West and larger ones could have occurred by chance only 27 times out of 10,000 samplings on the basis of the null hypothesis. Although this is a rather rare event, it is still not inconsistent with the null hypothesis. And since it is consistent with the null hypothesis, we necessarily run the risk of being wrong 27 times out of 10,000 if we rejected that theory. But when the probability drops to such a low level, we are wont to regard the observed difference as *statistically significant* rather than as a mere freak of sampling. We therefore state that we "reject the null hypothesis at the .0027 level of significance," since the probability of being wrong is so trifling.

It is somewhat analogous to a situation in which a student marks nine out of ten True–False questions correctly. With only one mistake out of a possible ten, could we assume that the student has merely guessed? Although this outcome would be possible by chance, the probability is not very high — approximately $\frac{1}{100}$. At this .01 significance level, most teachers would intuitively tend to reject the hypothesis that the student was totally ignorant, and that he had marked the answers in rather embrace? A liberal, humane teacher might prefer to run one type of risk, a rigid pedant would accept the other. The former would be disturbed by the possibility of failing a good student; the latter might shudder at the possibility of passing a poor one. Thus, two different observers might very well come to opposite decisions on identical statistical evidence because the cost of an incorrect decision will not weigh equally on each decision-maker. In fact, when there is "everything to gain and little or nothing to lose," or "everything to lose and nothing to gain," the statistical odds may even be completely ignored.

Such dilemmas are universal in everyday affairs. What are, for example, the consequences of condemning an innocent man, or releasing a guilty one? In this case, the moral standard of the culture usually dictates the decision. According to contemporary Anglo-Saxon ethics, it would be better to free a guilty man (accept false H) than run the risk of condemning an innocent one (reject true H). In some other cultures, the potential evil of allowing freedom to a heretic has been considered so great that the lesser evil has been to condemn the faithful innocent. The frustrated motorist cannot run the risk of parking alongside a fireplug which has not been used for 50 years because of the dire consequences of being wrong — that is, the expected loss in the remote case of fire.

Again, the probability of on-coming traffic around a blind curve may be very small; nevertheless, the prudent driver prefers the loss of a few minutes of time to the risk of life and limb; for the consequences of being wrong, however slight the probability, are so grave. However, in an emergency, even such risks are accepted, many times quite successfully.

From the foregoing analysis, it becomes evident that *statistics is able to measure the risk of being wrong, but it cannot advise an interested person whether or not to accept that risk.* Whether one accepts the measured risk will depend on considerations that are subjective, personal, ethical, and economic. Ironically enough, a statistical decision is thus grounded on non-statistical considerations. A penniless vagabond would avoid a given financial risk that would not deter a plunger with money in the bank; a robust man who abhors overshoes, raincoats, and umbrellas would accept the risk of rain which a consumptive would not dare to run.

There are, of course, many decisions which do not seem to matter either way, even when later events have proved them wrong. Either the costs of being wrong do not differ in seriousness, in which case the decision-maker may have a difficult dilemma on his hands; or the penalties of being wrong may not be too grievous. For example, in purely academic studies, such as whether there is a difference in personality scores between employed and unemployed wives, no policy or action may follow upon a statistical decision to reject the null hypothesis, which remains largely in the realm of paper work. Whatever consequences there were of being wrong may be indefinitely deferred, or in the end

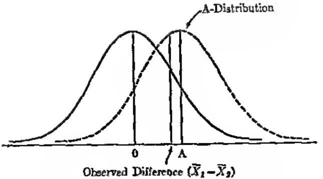


FIGURE 12.2 2 Sampling Distribution of Difference, Hypothetical Differences Between Population Means, 0 and A

as the risk of rejecting a true hypothesis decreases, the risk of accepting a false hypothesis correspondingly increases.* How, then, should we arbitrate between these two types of risks?

Assessment of Personal Risks. No reasonable man will ever decide an issue only on the basis of statistical probabilities. He will consider the uncomfortable consequences of being wrong as well as the agreeable benefits of deciding right. He will weigh these subjective alternatives as well as the objective quantitative probabilities. At this juncture, the modern statistician temporarily lays aside his statistics and falls back upon the pleasure–pain principle (the "felicific calculus") of nineteenth century utilitarianism.

In the foregoing cases, the teacher might have argued that it would be perfectly possible (even though not highly probable) for the student to have marked nine out of ten answers correctly by chance while knowing little or nothing about the subject. Being therefore still unconvinced about the student's knowledge, he could decide to fail him (acceptance of the null hypothesis that the observed sample is only a chance variation). To be sure, the teacher would thereby run the calculated risk of failing a good student who, in fact, was actually well informed. On the other hand, in passing the student, he would run the risk of passing a student who had, in fact, answered the questions *successfully by sheer* chance. Either decision will involve some risk of deciding wrong since *the null hypothesis can never be proved or disproved* anyway. In the light of personal or social consequences, which risk would the teacher

* In statistics, the rejection of a true hypothesis is termed a Type I, or alpha, error; while the acceptance of a false hypothesis is termed a Type II, or beta, error.

tion to analyze further the determining factors of the phenomenon unaer investigation — in our illustration, to determine whether the concept of "geographic region" should be probed further. Such a liberal interpretation of the critical ratios would create an opportunity for the enrichment of research, while rigid adherence to the stereotyped 1 and 5 per cent levels would prematurely cut off many possibilities of uncovering new variables. The implication is that we do not sufficiently exploit the uses of the null hypothesis when we terminate a study upon its rejection. The rejection of the hypothesis will provide an opportunity to continue study, and to investigate possible alternative hypotheses on the magnitude and nature of the difference which has been tentatively accepted as true. Unfortunately, many workers in social science do not avail themselves of the opportunity to take what is, after all, the next step in scientific inquiry. They are content to decide for or against the null hypothesis and go no further.

Other Applications of the Null Hypothesis

Comparison of Two Percentages. It is a common belief that women are more pacifistic than men. In measuring the degree of difference between sexes, let us suppose a sample survey in which 60 per cent of the men favor a militant foreign policy, but only 50 per cent of the women favor such a belligerent position. We display this result in a 2 × 2 table (Table 12.2.2). Is it possible, we may ask, that this difference of 10 percentage points is a mere sampling error, and that the percentage favoring a "get-tough" policy is actually the same in both populations? Such a difference could conceivably arise in the course of random sampling, even though the two sampled universes were exactly alike. Evidently, we have here another occasion for a test of the null hypothesis of no difference. A rejection of the null hypothesis would tend to sustain this popular belief, whereas acceptance would contradict it.

Table 12.2.2

Foreign Policy Favored, by Sex

Policy	Male	Female
Militant	60%	50%
Non-militant	40	50
	100%	100%
	$N = 400$	$N = 400$

Source Hypothetical

The technique of testing for the possible significance of the observed difference between two sample percentages is identical in fundamentals with that set forth for two sample means. To illustrate this procedure, even evaded. The decision is more in the nature of an intellectual adventure than a practical matter.

In any case, it should now be evident that decision-making is relativistic in character, differing from person to person and from group to group. An incorrect decision to accept what later turns out to be a defective carload of wheat would carry much more serious personal and commercial consequences than would an ascription of personality score differences to employed and unemployed wives where none actually exist.

The Level of Significance. If the foundation of the above relativistic arguments is reasonable, it would not seem very sound to adhere invariably to the .05 and .01 significance levels in all types of problems and situations, as is now the conventional practice. The conclusion seems inescapable that individuals, in their informal and unquantified dilemmas, often use more discernment in making their decisions than many statistical workers in the social sciences. If the field of statistics is an extension of common sense, this should not be so. There are, of course, occasions when high certainty might be deemed defensible. In the acceptance of a new drug, it would certainly seem humane to await the attainment of considerable certainty before concluding that the new drug differed significantly from the old. The over-eager floating of insufficiently tested medication might produce — and has, indeed, produced — deleterious effects on the population. But such is not the general attitude of the long-range gambler who can absorb non-fatal and short-term losses. His guiding rule is the old adage: "Nothing ventured, nothing gained" — a low threshold of risk. The more cautious and conservative decision rule is to "look before you leap" — a high threshold of risk. If we set the significance level too low — if we are overly cautious — we will exclude many new ideas and experiments; on the other hand, if we incautiously set it too high, we may wastefully follow many false leads.

Returning now to the difference between the suicide rates of eastern and western cities, we are actually interested in probing factors which cause the variation in rates. In such an exploratory study, it would be quite reasonable to enlarge the so-called *rejection zone* by lowering the critical ratio from 2.58 (1 per cent level) to 1.96 (5 per cent level) or even still lower. A tentative assumption of a genuine difference would even be justified at the one-sigma point (32 per cent level). The consequences of wrongly rejecting the "no-difference" hypothesis in this case are not very serious anyway; and the impulse to further investigation, which would have been suffocated by accepting the null hypothesis, might yield very pronounced benefits.

It should be recalled that the null hypothesis of no difference is, after all, merely a skirmish to determine whether there is sufficient justificathe stub of the table at 2.9 sigmas, we find that differences at least as large as ±2.9 can be expected to occur 2 times in 1,000. This completes the computational side of the problem, and puts us in a position to judge the plausibility of the null hypothesis.

In view of this small probability of being wrong, we might routinely reject the null hypothesis and tentatively accept the alternative hypothesis that women are more pacifistic than men. But even here, with only a $Pr = .002$, we would still have to consider the consequences of being wrong in that decision. Accordingly, we may imagine an anxious government spending millions of dollars in a time of crisis to reshape feminine opinion, on the assumption that women are pacifistic. But if that verdict is wrong — if women are already as militant as men — then millions will have been spent unnecessarily. For that reason the government might be reluctant to embark on an expensive propaganda campaign unless they were practically certain of the falsity of the null hypothesis. But insisting on practical certainty of its falsity has a price of its own: for, by reducing the risk of rejecting what could turn out to be a true hypothesis, we increase the risk of accepting a false one. This is the dilemma every decision-maker must somehow resolve. In this instance, it may appear less serious to reject the hypothesis of no difference (though true) than to accept it (though false). The propaganda can presumably do no harm, excepting the financial cost to an affluent government, and it may do some good.

The whole argument, which may seem needlessly prolix, may be wrapped up in the generalization that the null hypothesis will always be rejected whenever there is everything to gain if right, and nothing to lose if wrong. A person may asśume that he has no vitamin deficiency (H_0) and still take vitamin tablets (reject H_0) since he has everything to gain if he needs them and little or nothing to lose if he doesn't need them.

We reiterate that the detailed procedure set forth above for comparing two sample percentages is applicable only to large samples. Only then will the theoretical sampling distribution of differences be normal, as was assumed in setting up the probability of differences larger than the observed. Moreover, the samples must be extra large insofar as the sample percentages deviate markedly from 50 per cent. It is therefore necessary to scrutinize each application in order to ensure that the sample data meet the conditions which the technique presupposes, a rule which holds for every other statistical application as well.

Comparison of Three or More Percentages. There will be occasions to compare percentage distributions of samples from three or more universes in order to test the hypothesis that the universes do not differ — a typical null hypothesis. For example, the percentage of cross-class dating, as observed in samples of high school boys, may differ from one we process the samples of men and women shown in Table 12.2.2, assuming that each sample has 400 cases.

(1) Calculate the numerical difference between the sample percentages:

$$\begin{aligned} D &= 60 - 50 \\ &= 10 \end{aligned}$$

We conceive of this single observed difference as one of all possible differences within a sampling distribution which is assumed to be normal around a mean of zero as required by the null hypothesis.

(2) Estimate the standard error of the sampling distribution of the difference according to the formula:

$$\begin{aligned} s_D &= \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} \\ &= \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \end{aligned} \qquad (12.2.4)$$

where $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$, the weighted average of the two percentages, which estimates the common P, assumed by H_0

$q = 1 - p$

Solving:

$$\begin{aligned} p &= \frac{400(60) + 400(50)}{400 + 400} \\ &= 55 \end{aligned}$$

and:

$$\begin{aligned} s_D &= \sqrt{(55)(45)\left(\frac{1}{400} + \frac{1}{400}\right)} \\ &= \sqrt{\frac{4,950}{400}} \\ &= \sqrt{12.4} \\ &= 3.5 \end{aligned}$$

(3) Compute the significance ratio by dividing the estimated standard error into the observed difference, which reveals how many sigma units our observed difference is distant from the hypothetical zero mean:

$$\begin{aligned} z &= \frac{D}{s_D} \\ &= \frac{10}{3.5} \\ &= 2.9 \end{aligned}$$

(4) Enter the table of normal probabilities (Table I, Appendix) with this sigma measure, and fix the probability of obtaining a difference this large on the hypothesis that the true difference is zero. Entering

class, 27 (54 per cent) state that they engage in cross-class dating; 25 (50 per cent) of the middle class and 23 (46 per cent) of the lower class report this behavior. It is the differences among the sample frequencies that set up the problem: do these observed differences represent genuine differences among the three social classes? or are the populations probably homogeneous?

To provide an answer to this question, we *first* arrange the data in a 2 × 3 contingency table (Tahle 12.2.3b), which serves to display (a) the number of cases in each sample, (b) the number of cross-class daters in each sample, (c) the number of cross-class daters in all samples combined, and (d) the grand total of cases.

Table 12.2.3b *Social Class and Type of Dating, Observed Frequencies (O)*

Dating Type	Social Class			Total
	Upper	*Middle*	*Lower*	
Cross-Class	27	25	23	75
In-Class	23	25	27	75
Total	50	50	50	150

Second, we prepare a similar chart showing the expected frequency of each cell (Tahle 12.2.3c). These chance frequencies are derived according to the simple principle that percentage distributions within samples shall be identical with the corresponding marginal distribution. This marginal distrihution gives us the best estimate of the common percentage distribution posited by the null hypothesis. Since the dis-

Table 12.2.3c *Social Class and Type of Dating, Expected Frequencies (E)*

Dating Type	Social Class			Total
	Upper	*Middle*	*Lower*	
Cross-Class	25	25	25	75
In-Class	25	25	25	75
Total	50	50	50	150

social class to another (Table 12.2.3a), but it is still possible that these observed differences are a result of sampling variation rather than of social class practice.

Table 12.2.3a *Social Class and Type of Dating, Male High School Students*

Dating Type	Social Class		
	Upper	*Middle*	*Lower*
Cross-Class	54%	50%	46%
In-Class	46	50	54
	100%	100%	100%

Source Hypothetical

Chi-Square (χ^2) *as a Test of Significance.* When as many as three sample percentages are being compared, as in this example, a different technique is used from the one that applies to the case of only two samples. In order to compare all percentages simultaneously, it is the versatile *chi-square* technique that must now be employed. But the principles of testing remain the same: (1) the null hypothesis is formulated that social classes are indistinguishable in their dating behavior — that is, that the universe distributions are *homogeneous;* (2) a composite index (χ^2) of the magnitude of the observed differences is computed; which (3) permits us to fix the probability of obtaining larger percentage differences under the null hypothesis. Whether or not H_0 is rejected depends on the magnitude of that probability — a small probability points to its rejection, a large one to its acceptance.

The manner in which χ^2 measures the differences between the observed frequencies (percentages) and those expected under the null hypothesis has already been shown in the discussion of the contingency coefficient. As these differences increase in magnitude, the value of χ^2 likewise increases. Hence, the null hypothesis becomes progressively less tenable and more likely to be dismissed as χ^2 increases in magnitude. Thus, a relatively large χ^2 will tend to discredit the null hypothesis, whereas a small χ^2 will tend to uphold it, depending as usual on the calculated risks.

Computation of χ^2. Fundamentally, χ^2 is computed on the actual frequencies *instead of on* the percentages, since only in that manner will sample size be properly weighted. Reverting to Table 12.2.3a, let us suppose then that we have three samples of 50 boys each. Of the upper

Sixth, and finally, we sum all normed cell deviations to obtain χ^2. In symbols:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$
$$= .64$$

In our example, the sum of the normed deviations is .64. This is the end-statistic to which a probability (*Pr*-value) is attached for the purpose of testing the initial hypothesis of no difference between population proportions.

Fixing the Probability. In assessing the probability of a given χ^2-value, we cannot employ the familiar normal curve. Rather, we must turn to a different set of probabilities, which vary according to the number of degrees of freedom. Further, since it would be very impracticable to provide a full set of probabilities for each *df*, it is customary to present only several selected χ^2 significance values and their corresponding probabilities for each *df* up to 30. Beyond that point, the normal probabilities may be employed because of the convergence of χ^2 and *z*. Part of a conventional χ^2 table is presented in Table 12.2.4; it may be used to evaluate the significance of our computed χ^2-value of .64, 2 *df*. Examining the row entries alongside 2 *df* in the stub, we find that .64 is not significant at the 1 per cent level, nor is it significant at the 5 per cent level — or even at the 50 per cent level. In fact, better than 70 times in 100 we would obtain χ^2-values larger than .64 when the null bypothesis is true. In view of this result, we would in all likelihood hold

Table 12.2.4 *Table of χ^2-Values, by Selected Probability Values*

df	.70	.50	.30	.10	.05	.02	.01
1	.15	.46	1.07	2.71	3.84	5.41	6.64
2	.71	1.39	2.41	4.60	5.99	7.82	9.21
3	1.42	2.37	3.67	6.25	7.82	9.84	11.34
4	2.20	3.36	4.88	7.78	9.49	11.67	13.28
5	3.00	4.35	6.06	9.24	11.07	13.39	15.09
10	7.27	9.34	11.78	15.99	18.31	21.16	23.21
15	11.72	14.34	17.32	22.31	24.00	28.26	30.58
20	16.27	19.34	22.78	28.41	31.41	35.02	37.57
25	20.87	24.34	28.17	34.38	37.65	41.57	44.31
30	25.51	29.34	33.53	40.26	43.77	47.96	50.89

tribution of row totals is 50–50 per cent, we simply apply this ratio to each column sample of 50 items to obtain the expected frequency of 25 in each cell.

Degrees of Freedom (df). The χ^2 method requires that the marginal totals of the expected frequencies be identical with those of the observed cell frequencies; hence, in our problem it is possible to calculate freely only two of the six expected cell frequencies. After any two have been independently fixed, the other four are predetermined by the marginal totals and are therefore not free to vary. We therefore state that this table possesses "two degrees of freedom." In general, a "rows × columns" contingency table will have $(r-1)(c-1)$ degrees of freedom. This calculation is made necessary by the fact that the significance of the sample χ^2 is dependent in part on its degrees of freedom.

Continuing with the computation of χ^2, our *third* step is to subtract each expected frequency from its companion observed frequency (Table 12.2.3d). Any row (or column) of these deviations must sum to zero, a consequence of the aforementioned constraint that observed and expected marginal totals be equal.

Table 12.2.3d *Deviations of Observed from Expected Frequencies* $(O-E)$

Dating Type	Social Class			Total
	Upper	*Middle*	*Lower*	
Cross-Class	2	0	−2	0
In-Class	−2	0	2	0
Total	0	0	0	0

Fourth, we square the deviations — $(O-E)^2$ — and thereby eliminate signs:

4	0	4
4	0	4

Fifth, we divide each squared deviation by its expected frequency — $\frac{(O-E)^2}{E}$ — thereby norming it on its base:

.16	0	.16	.32
.16	0	.16	.32
.32	0	.32	$.64 = \chi^2$

The E-values are, of course, manipulated as in the previous demonstration, except that we here illustratively use the conventional compact worksheet (Table 12.2.6). With the obtained $\chi^2 = 45$, we consult the chi-square table to determine the probability of n value at least that large under H_0. Scanning the entries alongside $4[df = (r - 1)(c - 1)]$, in the stub, we discover that 45 is significant far beyond the 1 per cent level, since values even as large as 13.3 would occur only 1 time in 100. To determine how much less probable 45 is, it would be necessary to consult a table much more extensive than here available. We interpret this to mean that religion and occupation are not independent, but rather are linked together, since the observed pattern of joint frequencies is so unlikely under H_0.

Table 12.2.6 *Chi-Square Worksheet*

Cell No.	O	E	$(O - E)$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
11	20	10	10	100	10.0
12	5	10	−5	25	2.5
13	5	10	−5	25	2.5
21	5	10	−5	25	2.5
22	20	10	10	100	10.0
23	5	10	−5	25	2.5
31	5	10	−5	25	2.5
32	5	10	−5	25	2.5
33	20	10	10	100	10.0
			0		$\chi^2 = 45.0$

It should be repeated that the value, χ^2, is not a measure of the degree of association; its significance level only testifies to the probability of its presence. For a measure of the degree of correlation, a different technique would necessarily have to be resorted to. The reason for this is that raw chi-square is a variable quantity and not normed so as to yield an index ranging between the conventional limits of 0 and 1.

As with so many techniques employed in statistical testing, the chi-square method is applicable only under the circumstance that samples were randomly and independently selected. In addition, the chi-square table is valid only when each expected frequency is not small (minimum, 5–10). When these conditions are impossible of fulfillment, either the data must be adapted or alternative procedures must be employed.

to H_0; otherwise we would be confronted with a greater than 70 per cent risk of rejecting a true proposition, and this is obviously too great a risk to incur.

Chi-Square as a Test of Independence. Instead of focusing on sample differences as reflecting universe differences, it may be more useful to view the contingency table as reflecting the degree of association between the variables shown in the table. Thus, we may center our attention on the possible association between religion and occupation (Table 12.2.5)

Table 12.2.5 *Occupation and Religious Affiliation*

RELIGION	OCCUPATION			TOTAL
	Banker	*Merchant*	*Clerk*	
Presbyterian	20	5	5	30
Methodist	5	20	5	30
Baptist	5	5	20	30
TOTAL	30	30	30	90

Source. Hypothetical

and employ the chi-square test to establish the probability of any such association. In this instance, we again open with the null hypothesis that the two variables are statistically independent of each other. If we reject this hypothesis of independence, we tacitly accept the alternative that association is present, although the value of χ^2 does not supply us with a normed measure of the degree of that association.

The calculation of χ^2 for this problem is identical with that previously presented, except that we shall here employ a short-cut technique for computing the expected frequencies. To find the expected frequency of any given cell, we simply multiply its respective marginal totals, and divide this product by the grand total. Applying this rule to Table 12.2.5, we find the expected frequency of the first cell, first row, to be

$$E_{11} = \frac{30 \times 30}{90}$$
$$= 10$$

where the subscript "11" designates the intersection of the first row and first column. Once any four of the nine E-values have been freely computed in this manner, the remaining five may be calculated as residuals from the marginal totals which act as constraints.

mean of the sampling distribution is by hypothesis zero, we need only divide the observed r by the standard error:

$$\begin{aligned} z &= \frac{r-0}{\dfrac{1}{\sqrt{n-1}}} \\ &= r\sqrt{n-1} \end{aligned}$$

If a given z-measure should happen to be 2.58, the probability of larger r's on the assumption of the null hypothesis would be 1 in a 100, and we could reject it at the 1 per cent significance level; if the z-measure is 1.96, the probability of larger r's would be 5 in 100 and we could reject H_0 at the 5 per cent significance level. In practice, we do not need to carry out calculations of the above type; rather we consult a readily available table of critical values of r at varying levels of significance. An abbreviated table of that type is presented here by way of illustration. From Table 12.2.7, we see that an obtained r of .09 based on a sample of 500 cases would be significant at the 5 per cent level but not at the 1 per cent level. However, for a sample of only 50 cases, the obtained r must reach a value of .28 in order to be considered significant at the 5 per cent level. Thus, as sample size increases, we may reject the null hypothesis of zero correlation at a fixed significance level with a progressively smaller r-value.

Table 12.2.7 *Minimum Values of r, Selected Significance Levels and Sample Sizes*

SAMPLE SIZE (n)	LEVEL OF SIGNIFICANCE		
	5%	2%	1%
50	.28	.33	.36
60	.25	.30	.33
70	.23	.28	.31
80	.22	.26	.29
90	.21	.24	.27
100	.20	.23	.26
200	.14	.16	.18
300	.11	.13	.15
400	.10	.12	.13
500	.09	.10	.12

Null Hypothesis: Pearsonian r. The value of Pearsonian r is subject to sampling variation as is any other sample statistic. Thus, a first sample of observations on marriage adjustment and age difference may yield an r of .25, while a second yields an r of $-.25$. In fact, in a long succession of sample r's from this bivariate population, the correlation may now be plus, now minus, with neither sign displaying a tendency to dominate. Such a vacillation between signs would strongly suggest that the correlation in the sampled universe might be zero, and that each sample r was merely a result of sampling error ascribable to the chance factors operative in random sampling

As a rule, therefore, before rendering an elaborate interpretation of an obtained sample r, we should reckon with the possibility that there is really no linear relation at all. True, the exploration of social interaction would never be initiated for the express purpose of testing the null hypothesis of zero correlation — that is, to demonstrate the essential unrelatedness of things. Quite the contrary, the social analyst is moved by an interest in establishing functional relationships. Nevertheless, because of the tricks played by chance sampling, it is sound policy to investigate the assumption of independence between the two variables.

The testing for independence between two quantitative variables is in principle no different from that of qualitative variables, as set forth in the foregoing section: we allege the truth of the null hypothesis and then test that assumption by the obtained probability, or *Pr*-value. In this instance, we once again employ the normal probabilities, since the sampling distribution of r is approximately normal around the population correlation, with standard error expressed:

$$\sigma_r = \frac{(1 - r^2)}{\sqrt{n - 1}} \qquad (12.2.5)$$

provided that n is large, and the population correlation is not greater than about $\pm .8$. It follows that, when the population correlation is zero — as maintained by the null hypothesis — the sampling distribution will have a mean of zero and a standard error:

$$\sigma_r = \frac{1}{\sqrt{n - 1}} \quad . \qquad (12.2.6)$$

which is, of course, the relevant case for our purposes.

To test, then, the null hypothesis of no correlation, we conceive of the observed sample r as belonging to a normal distribution having a mean of zero and a standard error of $1/\sqrt{n - 1}$. To fix the probability of an r at least as large as that obtained, we transform the sample r to a sigma measure and find the corresponding probability. Since the *control* and *game theory*, which has tended to sharpen the edges of the idea and to inject a certain procedural clarification which it had not previously possessed. Then, too, with more extended dissemination of statistical tools, the technique has been gaining in general popularity. But the mere fact that it has fallen into the hands of a mass of new workers in the field of social studies, many of whom may not have absorbed a mature feeling for the decision-making process, has generated the type of ritualism which is here deplored.

The Meaning of "Hypothesis." In order to appreciate even more fully the meaning and purport of the null hypothesis, it will be helpful to distinguish the concept of hypothesis as employed by statisticians, and by their substantive sister sciences. Fundamentally, any hypothesis is a conceptual tool in the search for "truth." A tentative solution to a problem is provisionally set up, and "nature" is asked to confirm or reject it. If subsequent observation is congruent with the hypothesis, the hypothesis is accepted; if not, it is rejected, and another is set up until one is reasonably content with the fruitfulness of the inquiry.

However, the statistical use of hypotheses is distinguishable from that of the substantive sciences. In the latter area, the hypothesis is an explanatory element in research. Hypotheses of established prestige come readily to mind: the nebular hypothesis, the germ theory of disease, the molecular and atomic theories, the geocentric theory of the universe, the Malthusian theory of population growth, Ricardo's iron law of wages, Comte's law of three stages, Freud's Oedipus theory, Marx's economic determinism — all were launched with considerable confidence that they would be confirmed. They were rational, plausible explanations of events, on the basis of which the past was rendered credible and the future predictable. They were thought to be "true."

A statistical hypothesis, on the other hand, is not primarily set up for explanatory purposes, although an explanation may be implied. Fundamentally, a statistical hypothesis is a description of a population characteristic, the plausibility of which is checked against the evidence in the given sample. This statistically testable hypothesis of a parameter value is the null hypothesis. We (1) set up a hypothesis of the parameter value, and (2) determine whether we could reasonably expect the observed value to have been obtained by chance sampling variation on the assumption of the proclaimed hypothesis.

As implied in the preceding presentations, the most prevalent use of the null hypothesis, however, has arisen in what we may call the experimental situation; and it is from this context that its predominant connotation in current usage has emerged. We have here another instance of a generic concept being appropriated by a specific subcategory. The hypothesis, as now usually employed, may take on several related formuthe concepts "spurious," "concealed factor," "selective factor," and the like beset the social analyst whenever he ottempts to translate an inanimate numerical statement into a living guide to action. In our problem of suicide, it will not be obvious on the surface, for example, whether it is "eastness" and "westness" that determine the observed difference between suicide rates, or whether there are unidentified and concealed factors that are doing the work falsely credited to "regional" factors. Similarly, we cannot be sure that it is habitual smoking that determines the observed difference in the incidence of lung cancer between smokers and non-smokers. And even in experimentation, we may spuriously credit the experimental factor with work being performed hy unknown, hidden variables.

Students who have not firmly gripped the fundamentals of decision-making are not likely to appreciate its probabilistic and inconclusive nature. For example, "no proof of a difference" may be mistaken for "proof of no difference." Similarly, the null hypothesis may be "accepted" rather than merely "not rejected." It cannot be too sternly stated: the null hypothesis can never be proved, and in this strict sense can never be accepted. A more precise, though more cumbersome, statement would be: "not to reject the null hypothesis at a given level of confidence."

There is something to be said for the point of view that, a priori, it is extremely unlikely that there should be any absolute identity between the parameters of populations in the first place, no matter how strong the evidence is for no difference. In practice, the conception of no difference is probably informally and generally interpreted as "no appreciable difference." In fact, any observed difference is more consistent with an alternative hypothesis of a specific difference than with the null hypothesis of no difference. After all, the observed difference is the best estimate of the true difference. This being the case, we should give as much consideration to an alternative hypothesis of some specific difference as to the null hypothesis of no difference.

QUESTIONS AND PROBLEMS

1. Define the following concepts:
   Hypothesis Testing
   Statistical Hypothesis
   Null Hypothesis
   Decision-Making
   Significance (Critical) Ratio
   Sampling Distribution of the Difference
   Chi-Square
   Test of Independence
   Test of Homogeneity

lations: "no difference," "zero correlation," "chance outcome," "random variation." In an experiment we seek to determine whether an experimental group, having been given a treatment, differs from the control group, thereby establishing the probable effect of the experimental factor. In order to proceed in the most parsimonious manner, it is provisionally assumed that there is no real difference at all, and that any observed difference is only a chance variation around a zero difference. If a difference should be established with a high enough probability, we should ideally turn our attention to an alternative hypothesis about the possible magnitude of that difference. Unless we know the approximate magnitude of the difference, we are often not much better off for purposes of subsequent action than we were before we accepted the mere presence of some difference. Accordingly, it has been suggested that a confidence interval for an observed difference will often be more useful than the critical ratio which tests the null hypothesis.

The "Social Experiment." But the concept of "treatment" as employed in biological studies, for example, must be liberally interpreted in sociological investigations. Although experimental procedures are by no means unknown, they are not widely used, since the human animal is not so readily accessible to experimentation. Even though experimental methods may be considered by some methodologists as the only valid ones, we certainly cannot insist on such methodological purity in the present state of sociological science. For some time to come we will have to rely on "Nature" to produce crimes, divorces, births, and deaths under diverse and uncontrolled conditions which can be standardized and held constant only with some difficulty. In practice, we therefore frequently make survey studies, or observations of "nature in the raw," with no factors under laboratory control. Nevertheless, we still find it profitable to apply the null hypothesis in such contexts, sometimes disparagingly termed *ex post facto* investigations, as well as under conditions that are more favorable to randomization, as required by the classical experiment. There is some disagreement among students of methodology concerning the relative values of such procedures.*

Interpretive Issues in Decision-Making. The decisions consequent upon the null hypothesis are encumbered with dilemmas very similar to those that plague the interpretation of a correlation. Nature is very complex, and any formula will accommodate only a limited number of variables, while at the same time responding to innumerable hidden factors. Hence,

* Cf. Hanan C. Selvin, "A Critique of Tests of Significance in Survey Research," *American Sociological Review*, XXII, 1957, pp. 519–527; Robert McGinnis, "Randomization and Inference in Sociological Research," *American Sociological Review*, XXIII, 1958, pp. 408–414; Leslie Kish, "Some Statistical Problems in Research Design," *American Sociological Review*, XXIV, 1959, pp. 328–338.

(a) Compute χ^2.

(b) Is the difference significant at the .05 level? at the .02 level?

(c) What substantive conclusion would you draw?

(d) Is this borne out by direct analysis of the table?

Selected References

Braithwaite, Richard B., *Scientific Explanation.* Cambridge University Press, New York, 1957. Chapters 5 and 7.

Fisher, Ronald A., *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1942. Chapter 1.

Neyman, Jerzy, *First Course in Probability and Statistics.* Henry Holt and Company, New York, 1950. Chapter 5.

Savage, Leonard J., *The Foundations of Statistics.* John Wiley & Sons, Inc., New York, 1954. Chapter 4.

Wallis, W. Allen, and Harry V. Roberts, *Statistics: A New Approach.* The Free Press, Glencoe, Illinois, 1956. Pages 395–396.

2. Draw two samples of 30 cases each from the list of 107 suicide rates (Table 3.1.1a). Compute sample means and proceed to test the null hypothesis of no difference. Would it be possible to reject the null hypothesis, even though both samples were drawn from the same population?

3. Of 225 college men in a sample survey, 45 per cent favor the abolition of fraternities; of 100 college women, 40 per cent favor such a policy. Test the null hypothesis that men and women do not differ in their attitudes on this issue.

4. A candidate claimed that 60 per cent of the electorate would vote for him. In a sample of 1,000 registered voters, 55 per cent declared for that candidate. By means of chi-square, test the credibility of the candidate's claim.

5. An investigator compared questionnaire and interview responses to a given statement and obtained the results shown in Table 12.2.8. Use the chi-square test to determine whether the difference between responses is significant at the 5 per cent level.

Table 12.2.8

Comparison of Questionnaire and Interview Results, 69 Subjects

Response Category	Interview	Questionnaire
1. Strongly agree	37	25
2. Agree	17	27
3. Indifferent	14	10
4. Disagree	1	7
5. Strongly disagree	0	0
	69	69

6. Table 12.2.9 represents an attempt to measure the impact of the Kinsey findings on tolerance toward sexual practices.

Table 12.2.9 *Impact of Kinsey Findings, Sex Tolerance Before and After Exposure, 1952*

Groups	Point Change in Direction of Tolerance				Total
	0	1	2	3 & over	
Experimental (with Kinsey data)	32	27	21	24	104
Control (without Kinsey data)	52	21	21	12	106
					210

Source: C. Kirkpatrick, S. Stryker, and P. Buell, "Attitudes Toward Male Sex Behavior," *American Sociological Review*, XVII, 1952, p. 580.

Appendix

Table II *Ordinates of the Normal Curve*

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	1.00000	.99995	.99980	.99955	.99920	.99875	.99820	.99755	.99685	.99596
0.1	.99501	.99396	.99283	.99158	.99025	.98881	.98728	.98565	.98393	.98211
0.2	.98020	.97819	.97609	.97390	.97161	.96923	.96676	.96420	.96156	.95882
0.3	.95600	.95309	.95010	.94702	.94387	.94055	.93723	.93382	.93024	.92677
0.4	.92312	.91399	.91558	.91169	.90774	.90371	.89961	.89543	.89119	.88688
0.5	.88250	.87805	.87353	.86896	.86432	.85962	.85488	.85006	.84519	.84060
0 6	.83527	.83023	.83514	.82010	.81481	.80957	.80429	.79896	.79459	.78817
0.7	.78270	.77721	.77167	.76610	.76048	.75484	.74916	.74342	.73769	.73193
0.8	.72615	.72033	.71448	.70861	.70272	.69681	.69087	.68493	.67896	.67298
0.9	.66689	.66097	.65494	.64891	.64287	.63683	.63077	.62472	.61865	.61259
1.0	.60653	.60047	.59440	.58834	.58228	.57623	.57017	.56414	.55810	.55209
1.1	.54607	.54007	.53409	.52812	.52214	.51620	.51027	.50437	.49848	.49260
1.2	.48675	.48092	.47511	.46933	.46357	.45793	.45212	.44644	.44078	.43516
1.3	.42956	.42399	.41845	.41294	.40747	.40202	.39661	.39123	.38569	.38058
1.4	.37531	.37007	.36487	.35971	.35459	.34950	.34445	.33944	.33447	.32954
1.5	.32465	.31980	.31500	.31023	.30550	.30082	.29618	.29158	.28702	.28251
1.6	.27804	.27361	.26923	.26489	.26059	.25634	.25213	.24797	.24385	.23978
1.7	.23575	.23176	.22782	.22392	.22008	.21627	.21251	.20879	.20511	.20148
1.8	.19790	.19436	.19086	.18741	.18400	.18064	.17732	.17404	.17081	.16762
1.9	.16448	.16137	.15831	.15530	.15232	.14939	.14650	.14364	.14083	.13806
2.0	.13534	.13265	.13000	.12740	.12483	.12230	.11981	.11737	.11496	.11259
2.1	.11025	.10795	.10570	.10347	.10129	.09914	.09702	.09495	.09290	.09090
2.2	.08892	.08698	.08507	.08320	.08136	.07956	.07773	.07604	.07433	.07265
2.3	.07100	.06939	.06780	.06624	.06471	.06321	.06174	.06029	.05888	.05750
2.4	.05614	.05481	.05350	.05222	.05096	.04973	.04852	.04737	.04618	.04505
2.5	.04394	.04285	.04179	.04074	.03972	.03873	.03775	.03680	.03586	.03494
2.6	.03405	.03317	.03232	.03148	.03066	.02986	.02908	.02831	.02757	.02684
2.7	.02612	.02542	.02474	.02408	.02343	.02280	.02218	.02157	.02098	.02040
2.8	.01984	.01929	.01876	.01823	.01772	.01723	.01674	.01627	.01581	.01536
2.9	.01497	.01449	.01408	.01367	.01328	.01288	.01252	.01215	.01179	.01145
3 0	.01111									
4 0	.00034									

Source: Herbert Arkin and Raymond R. Colton, *Tables for Statisticians*, Barnes and Noble, Inc., New York, 1950, p. 115, Table 11.

Table I *Areas of the Normal Curve*

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0 0	.0000	0040	.0080	.0120	.0159	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0 6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0 8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0 9	3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3718	.3729	.3749	.3770	.3790	.3810	.3830
1 2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4083	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.6	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4430	.4441
1.6	.4452	.4463	.4474	.4485	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4590	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4758	.4762	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4865	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2 5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4980	.4980	.4981
2.9	.4981	.4982	.4983	.4984	.4984	.4984	.4985	.4985	.4986	.4986
3 0	.49865	.4987	.4987	.4988	.4988	.4988	.4989	.4989	.4989	.4990
3 1	.49903	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
4.0	.49997									

Source: Herbert Arkin and Raymond R Colton, *Tables for Statisticians*, Barnes and Noble, Inc., New York, 1950, p. 114, Table 10.

Table III *Values of* χ^2 *(Continued)*

df	*P* = .30	.20	.10	.05	.02	.01	.001
1	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	3.665	4.642	6.251	7.815	9.837	11.345	16.268
4	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	6 064	7.289	9.236	11.070	13.388	15.086	20.517
6	7.231	8.558	10.645	12.592	15.033	16.812	22 457
7	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	14 011	15.812	18.549	21.026	24 054	26.217	32.909
13	16.119	16.985	19.812	22.362	25.472	27.688	34.528
14	16.222	18.151	21.064	23.685	26 873	29.141	36.123
15	17.322	19.311	22.307	24.996	28.259	30.678	37.697
16	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	22.775	25.038	28.412	31.410	35.020	37.566	45.316
21	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	29.246	31.795	35.563	38.885	42.856	45 642	54.052
27	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	31.391	34.027	37.916	41.337	45.419	48 278	56 893
29	32.461	35.139	39.087	42 557	46.693	49.588	58 302
30	33.530	36.250	40.256	43.773	47.962	50.892	59.703

For larger values of *df*, the expression $\sqrt{2\chi^2} - \sqrt{2df - 1}$ may be used as a normal deviate with unit variance, remembering that the probability of χ^2 corresponds with that of a single tail of the normal curve.

Table III *Values of* χ^2

df	*P* = .99	.98	.95	.90	.80	.70	.50
1	.000157	.000628	.00393	.0158	.0642	.148	.455
2	.0201	.0404	.103	.211	.446	.713	1.386
3	.115	.185	.352	.584	1.005	1.424	2.366
4	.297	.429	.711	1.064	1.649	2.105	3.357
5	.554	.752	1.145	1.610	2.343	3.000	4.351
6	.872	1.134	1.635	2.204	3.070	3.828	5 348
7	1.239	1 564	2 167	2.833	3.822	4.671	6 346
8	1.646	2 032	2.733	3.490	4.594	5 527	7.344
9	2 088	2.532	3.325	4.168	5.380	6.393	8 343
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342
11	3 053	3 609	4.575	5.578	6.989	8.148	10.341
12	3 571	4.178	5.226	6 304	7.807	9 034	11.340
13	4.107	4.765	5 892	7.042	8.634	0.026	12.340
14	4 660	5 368	6.571	7.790	9 467	10.821	13 339
15	5 229	5.985	7.261	8.547	10 307	11.721	14.339
16	5 812	6 614	7.962	9.312	11.152	12 624	15 338
17	6.408	7.255	8 672	10 085	12 002	13.531	16.338
18	7 015	7.906	9 390	10.865	12.857	14.440	17.338
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338
20	8 260	9.237	10.851	12.443	14.578	16.266	19 337
21	8 897	9.015	11.591	13 240	15.445	17.182	20 337
22	0 642	10.600	12 338	14.041	16.314	18.101	21.337
23	10.196	11.293	13 091	14 848	17.187	10 021	22.337
24	10 856	11.992	13 848	15 659	18 062	19 943	23.337
25	11.524	12 697	14.611	16.473	18 940	20 867	24 337
26	12.198	13.409	15 379	17.292	19 820	21.792	25.336
27	12 879	14.125	16.151	18.114	20.703	22.719	26 336
28	13 565	14.847	16 928	18.939	21.588	23.647	27.336
29	14 256	15.574	17.708	19.768	22 475	24.577	28 336
30	14.953	16.306	18.493	20.598	23 364	25 508	29 336

Source: Reprinted from Table IV of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, published by Oliver & Boyd Ltd., Edinburgh, and by permission of the authors and publishers.

Table IV *Five Thousand Random Digits (Continued)*

	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99
50	32847	31282	03345	80593	69214	70381	78285	20054	91018	16742
51	16916	00041	30236	55023	14253	76582	12092	86533	92426	37655
52	66176	34047	21005	27137	03191	48970	64625	22394	39622	79085
53	46299	13335	12180	16861	38043	69292	62675	63631	37020	78195
54	22847	47839	45385	23289	47526	54098	45683	55849	51575	64689
55	41851	54160	92320	69936	34803	92479	33399	71160	64777	83378
56	28444	59497	91586	95917	68553	28639	06455	34174	11130	91994
57	47520	62378	98855	83174	13088	16561	68559	26679	06238	51254
58	34978	63271	13142	82681	05271	08822	06490	44984	49307	62717
59	37404	80416	69035	92980	49486	74378	75610	74976	70056	15478
60	32400	65482	52099	53676	74648	94148	65095	69597	52771	71551
61	89262	86332	51718	70663	11623	29834	79820	73002	84886	03591
62	86866	09127	98021	03871	27789	58444	44832	36505	40672	30180
63	90814	14833	08759	74645	05046	94056	99094	65091	32663	73040
64	19192	82756	20553	58446	55376	88914	75096	26119	83898	43816
65	77585	52593	56612	95766	10019	29531	73064	20953	53523	58136
66	23757	16364	05096	03192	62386	45389	85332	18877	55710	96459
67	45989	96257	23850	26216	23309	21526	07425	50254	19455	29315
68	92970	94243	07316	41467	64937	52406	25225	51553	31220	14032
69	74346	59596	40088	98176	17896	86900	20249	77753	19099	48885
70	87646	41309	27636	45153	29988	94770	07255	70908	05340	99751
71	80099	71038	45146	06146	55211	99429	43169	66259	97786	59180
72	10127	46900	64984	75348	04115	33624	68774	60013	35515	62556
73	67995	61977	18984	64091	02785	27762	42529	97144	80407	64524
74	26304	80217	84934	82657	69291	35397	98714	35104	08187	48109
75	81994	41070	56642	64091	31229	02595	13513	45148	78722	80144
76	59537	34662	79631	89403	65212	09975	06118	86197	58208	16162
77	51228	10937	62396	81460	47331	91403	95007	06047	16846	64809
78	31089	37995	29577	07828	42272	54016	21950	86192	99046	84864
79	38207	97938	93459	75174	79460	55436	67206	87644	21296	43395
80	88666	31142	09474	89712	63153	62333	42212	06140	42594	43671
81	53365	56134	67582	92557	89520	33452	05134	70628	27612	33738
82	89807	74530	38004	90102	11693	90257	05500	79920	62700	43325
83	18682	81038	85662	90915	91631	22223	91588	80774	07716	12548
84	63571	32579	63942	25371	09234	94592	98475	76884	37635	33608
85	68927	56492	67799	95398	77642	54913	91853	08424	81450	76229
86	56401	63186	39389	88798	31356	89235	97036	32341	33292	73757
87	24333	95603	02359	72942	46287	95382	08452	62862	97869	71775
88	17025	84202	95199	62272	06366	16175	97577	99304	41587	03686
89	02804	08253	52133	20224	68034	50865	57868	22343	55111	03607
90	08298	03879	20995	19850	73090	13191	18963	82244	78479	99121
91	59883	01785	82403	96062	03785	03488	12970	64896	38336	30030
92	46982	06682	62864	91837	74021	89094	39952	64158	79614	78235
93	31121	47266	07661	02051	67599	24471	69843	83696	71402	76287
94	97867	56641	63416	17577	30161	87320	37752	73276	48969	41915
95	57364	86746	08415	14621	49430	22311	15836	72492	49372	44103
96	09559	26263	69511	28064	75999	44540	13337	10918	79846	54809
97	53873	55571	00608	42661	91332	63956	74087	59008	47493	99581
98	35531	19162	86406	05299	77511	24311	57257	22826	77555	05941
99	28229	88629	25695	94932	30721	16197	78742	34974	97528	45447

Table IV *Five Thousand Random Digits*

	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99
00	59391	58030	52098	82718	87024	82848	04190	96574	90464	29065
01	99567	76364	77204	04615	27062	96621	43918	01896	83991	51141
02	10363	97518	51400	25670	98342	61891	27101	37855	06235	33316
03	86859	19558	64432	16706	99612	59798	32803	67708	15297	28612
04	11258	24591	36863	55368	31721	94335	34936	02566	80972	08188
05	95068	88628	35911	14530	33020	80428	39936	31855	34334	64865
06	54463	47737	73800	91017	36239	71824	83671	39892	60518	37092
07	16874	62677	57412	13215	31389	62233	80827	73917	82802	84420
08	92494	63157	76593	91316	03505	72389	96363	52887	01087	66091
09	15669	56689	35682	40844	53256	81872	35213	09840	34471	74441
10	99116	75486	84989	23476	52967	67104	39495	39100	17217	74073
11	15696	10703	65178	90637	63110	17622	53988	71087	84148	11670
12	97720	15369	51269	69620	03388	13699	33423	67453	43269	56720
13	11666	13841	71681	98000	35979	39719	81899	07449	47985	46967
14	71628	73130	78783	75691	41632	09847	61547	18707	85489	69944
15	40501	61089	99943	91843	41995	88931	73631	69361	05375	15417
16	22518	55576	98215	82068	10798	86211	36584	67466	69373	40054
17	75112	30465	62173	02132	14878	92879	22281	16783	86352	00077
18	80327	02671	98191	84342	90813	49268	95441	15496	20168	09271
19	60251	45548	02146	05597	48228	81366	34598	72856	66762	17002
20	57430	82270	10421	05540	43648	75888	66049	21511	47676	33444
21	73528	39559	34434	88596	54076	71693	43132	14414	79949	85193
22	25991	65959	70769	64721	86413	33475	42740	06175	82758	66248
23	78388	16638	09134	59880	63806	48472	39318	35434	24057	74739
24	12477	09965	96657	57994	59439	76330	24596	77515	09577	91871
25	83266	32883	42451	15579	38155	29793	40914	65990	16255	17777
26	76970	80876	10237	39515	79152	74798	39357	09054	73579	92359
27	37074	65198	44785	68624	98336	84481	97610	78735	46703	98265
28	83712	06514	30101	78295	54656	85417	43189	60048	72781	72606
29	20287	56862	69727	94443	64936	08366	27227	05158	50326	59566
30	74261	32592	86538	27041	65172	85532	07571	80609	39285	65340
31	64081	49863	08478	96001	18888	14810	70545	89755	59064	07210
32	05617	75818	47750	67814	29575	10526	66192	44464	27058	40467
33	26793	74951	95466	74307	13330	42664	85515	20632	05497	33625
34	65988	77850	48737	54719	52056	01596	03845	35067	03134	70322
35	27366	42271	44300	73399	21105	03280	73457	43093	05192	48657
36	56760	10909	98147	34736	33863	05256	12731	66598	50771	83665
37	72880	43338	93643	58904	59543	23943	11231	83268	65938	81581
38	77888	38100	03062	58103	47961	83841	25878	23746	55903	44115
39	28440	07819	21580	51459	47971	29882	13990	29226	23608	15873
40	63525	94441	77033	12147	51054	49955	58312	76923	96071	05813
41	47606	93410	16359	89033	89696	47731	64498	31776	05383	39902
42	52669	45030	96279	14709	52372	87832	02735	50803	72744	88208
43	16738	60159	07425	62369	07515	82721	37875	71153	21315	00132
44	59348	11695	45751	15865	74739	05572	32688	20271	65128	14551
45	12900	71775	29845	60774	94924	21810	38636	33717	67598	82521
46	75086	23537	49939	33595	13484	97588	28617	17979	70749	35234
47	99495	51434	29181	09993	38190	42553	68922	52125	91077	40197
48	26075	31671	45386	36583	93459	48599	52022	41330	60651	91321
49	13636	93596	23377	51133	95126	61496	42474	45141	46660	42338

Source: Abridged from Table 1.5.1 in G. W. Snedecor, *Statistical Methods*, 5th edition, Iowa State University Press, Ames, Iowa, 1956, *by permission of the publisher.*

Table V *Squares and Square Roots of the Numbers from 1 to 1,000* *(Continued)*

Number	*Square*	*Square Root*	*Number*	*Square*	*Square Root*
101	1 02 01	10.050	151	2 28 01	12.288
102	1 04 04	10.100	152	2 31 04	12.329
103	1 06 09	10.149	153	2 34 09	12.369
104	1 08 16	10.198	154	2 37 16	12.410
105	1 10 25	10.247	155	2 40 25	12.450
106	1 12 36	10.296	156	2 43 36	12.490
107	1 14 49	10.344	157	2 46 49	12.530
108	1 16 64	10.392	158	2 49 64	12.570
109	1 18 81	10.440	159	2 52 81	12.610
110	1 21 00	10.488	160	2 56 00	12.649
111	1 23 21	10.536	161	2 59 21	12.689
112	1 25 44	10.583	162	2 62 44	12.728
113	1 27 69	10.630	163	2 65 69	12.767
114	1 29 96	10.677	164	2 68 96	12.806
115	1 32 25	10.724	165	2 72 25	12.845
116	1 34 56	10.770	166	2 75 56	12.884
117	1 36 89	10.817	167	2 78 89	12.923
118	1 39 24	10.863	168	2 82 24	12.961
119	1 41 61	10.909	169	2 85 61	13.000
120	1 44 00	10.954	170	2 89 00	13.038
121	1 46 41	11.000	171	2 92 41	13.077
122	1 48 84	11.045	172	2 95 84	13.115
123	1 51 29	11.091	173	2 99 29	13.153
124	1 53 76	11.136	174	3 02 76	13.191
125	1 56 25	11.180	175	3 06 25	13.229
126	1 58 76	11.225	176	3 09 76	13.266
127	1 61 29	11.269	177	3 13 29	13.304
128	1 63 84	11.314	178	3 16 84	13.342
129	1 66 41	11.358	179	3 20 41	13.379
130	1 69 00	11.402	180	3 24 00	13.416
131	1 71 61	11.446	181	3 27 61	13.454
132	1 74 24	11.489	182	3 31 24	13.491
133	1 76 89	11.533	183	3 34 89	13.528
134	1 79 56	11.576	184	3 38 56	13.565
135	1 82 25	11.619	185	3 42 25	13.601
136	1 84 96	11.662	186	3 45 96	13.638
137	1 87 69	11.705	187	3 49 69	13.675
138	1 90 44	11.747	188	3 53 44	13.711
139	1 93 21	11.790	189	3 57 21	13.748
140	1 96 00	11.832	190	3 61 00	13.784
141	1 98 81	11.874	191	3 64 81	13.820
142	2 01 64	11.916	192	3 68 64	13.856
143	2 04 49	11.958	193	3 72 49	13.892
144	2 07 36	12.000	194	3 76 36	13.928
145	2 10 25	12.042	195	3 80 25	13.964
146	2 13 16	12.083	196	3 84 16	14.000
147	2 16 09	12.124	197	3 88 09	14.036
148	2 19 04	12.166	198	3 92 04	14.071
149	2 22 01	12.207	199	3 96 01	14.107
150	2 25 00	12.247	200	4 00 00	14.142

Table V *Squares and Square Roots of the Numbers from 1 to 1,000*

Number	*Square*	*Square Root*	*Number*	*Square*	*Square Root*
1	1	1.000	51	26 01	7.141
2	4	1.414	52	27 04	7.211
3	9	1.732	53	28 09	7.280
4	16	2.000	54	29 16	7.348
5	25	2.236	55	30 25	7.416
6	36	2.449	56	31 36	7.483
7	49	2.646	57	32 49	7.550
8	64	2.828	58	33 64	7.616
9	81	3.000	59	34 81	7.681
10	1 00	3.162	60	36 00	7.746
11	1 21	3.317	61	37 21	7.810
12	1 44	3.464	62	38 44	7.874
13	1 69	3.606	63	39 69	7.937
14	1 96	3.742	64	40 96	8.000
15	2 25	3.873	65	42 25	8.062
16	2 56	4.000	66	43 56	8.124
17	2 89	4.123	67	44 89	8.185
18	3 24	4.243	68	46 24	8.246
19	3 61	4.359	69	47 61	8.307
20	4 00	4.472	70	49 00	8.367
21	4 41	4.583	71	50 41	8.426
22	4 84	4.690	72	51 84	8.485
23	5 29	4.796	73	53 29	8.544
24	5 76	4.899	74	54 76	8.602
25	6 25	5.000	75	56 25	8.660
26	6 76	5.099	76	57 76	8.718
27	7 29	5.196	77	59 29	8.775
28	7 84	5.292	78	60 84	8.832
29	8 41	5.385	79	62 41	8.888
30	9 00	5.477	80	64 00	8.944
31	9 61	5.568	81	65 61	9.000
32	10 24	5.657	82	67 24	9.055
33	10 89	5.745	83	68 89	9.110
34	11 56	5.831	64	70 56	9.185
35	12 25	5.916	85	72 25	9.220
36	12 96	6.000	86	73 96	9.274
37	13 69	6.083	87	75 69	9.327
38	14 44	6.164	88	77 44	9.381
39	15 21	6.245	89	79 21	9.434
40	16 00	6.325	90	81 00	9.487
41	16 81	6.403	91	82 81	9.539
42	17 64	6.481	92	84 64	9.592
43	18 49	6.557	93	86 49	9.644
44	19 36	6.633	94	88 36	9.695
45	20 25	6.708	95	90 25	9.747
46	21 16	6.782	96	92 16	9.798
47	22 09	6.856	97	94 09	9.849
48	23 04	6.928	98	96 04	9.899
49	24 01	7.000	99	98 01	9.950
50	25 00	7.071	100	1 00 00	10.000

Table V *Squares and Square Roots of the Numbers from 1 to 1,000.* *(Continued)*

Number	*Square*	*Square Root*
301	9 06 01	17.349
302	9 12 04	17.378
303	9 18 09	17.407
304	9 24 16	17.436
305	9 30 25	17.464
306	9 36 36	17.493
307	9 42 49	17.521
308	9 48 64	17.550
309	9 54 81	17.578
310	9 61 00	17.607
311	9 67 21	17.635
312	9 73 44	17.664
313	9 79 69	17.692
314	9 85 96	17.720
315	9 92 25	17.748
316	9 98 56	17.776
317	10 04 89	17.804
318	10 11 24	17.833
319	10 17 61	17.661
320	10 24 00	17.889
321	10 80 41	17.916
322	10 36 84	17.944
323	10 43 29	17.972
324	10 49 76	18.000
325	10 56 25	18.028
326	10 62 76	18.055
327	10 69 29	18.083
328	10 75 84	18.111
329	10 82 41	18.138
330	10 89 00	18.166
331	10 95 61	18.193
332	11 02 24	18.221
333	11 08 89	18.248
334	11 15 56	18.278
335	11 22 25	18.303
336	11 28 96	18.330
337	11 35 69	18.358
338	11 42 44	18.385
339	11 49 21	18.412
340	11 56 00	18.439
341	11 62 81	18.466
342	11 69 64	18.493
343	11 76 49	18.520
344	11 83 36	18.547
345	11 90 25	18.574
346	11 97 16	18.601
347	12 04 09	18 628
348	12 11 04	18.655
349	12 18 01	18.682
350	12 25 00	18.708
351	12 32 01	18.735
352	12 39 04	18.762
353	12 46 09	18 788
354	12 53 16	18.815
355	12 60 25	18.841
356	12 67 36	18.868
357	12 74 49	18.894
358	12 81 64	18.921
359	12 88 81	18.947
360	12 96 00	18.974
361	13 03 21	19.000
362	13 10 44	19.026
363	13 17 69	19.053
364	13 24 96	19.079
365	13 32 25	19.105
366	13 39 56	19.131
367	13 46 89	19.157
368	13 54 24	19.183
369	13 61 61	19.209
370	13 69 00	19.235
371	13 76 41	19.261
372	13 83 84	19.287
373	13 91 29	19.313
374	13 98 76	19 339
375	14 06 25	19.363
376	14 18 76	19.391
377	14 21 29	19.416
378	14 28 84	19.442
379	14 36 41	19.468
380	14 44 00	19.494
381	14 51 61	19 519
382	14 59 24	19 545
383	14 66 89	19.570
384	14 74 56	19.596
385	14 82 25	19.621
386	14 89 96	19 647
387	14 97 69	19.672
388	15 05 44	19.698
389	15 13 21	19.723
390	15 21 00	19.748
391	15 28 81	19.774
392	15 36 64	19.799
393	15 44 49	19.824
394	15 52 36	19.849
395	15 60 25	19.875
396	15 68 16	19.900
397	15 76 09	19.925
398	15 84 04	19 950
399	15 92 01	19.975
400	16 00 00	20.000

Squares and Square Roots of the Numbers from 1 to 1,000

Table V (*Continued*)

Number	*Square*	*Square Root*
201	4 04 01	14.177
202	4 08 04	14.213
203	4 12 09	14.248
204	4 16 16	14.283
205	4 20 25	14.318
206	4 24 36	14 353
207	4 28 49	14.387
208	4 32 64	14.422
209	4 36 81	14.457
210	4 41 00	14.491
211	4 45 21	14.526
212	4 49 44	14 560
213	4 53 69	14 595
214	4 57 96	14.629
215	4 62 25	14 663
216	4 66 56	14 697
217	4 70 89	14.731
218	4 75 24	14 765
219	4 79 61	14.799
220	4 84 00	14.832
221	4 88 41	14 866
222	4 92 84	14.900
223	4 97 29	14.933
224	5 01 76	14.967
225	5 06 25	15 000
226	5 10 76	15 033
227	5 15 29	15 067
228	5 19 84	15.100
229	5 24 41	15.133
230	5 29 00	15.166
231	5 33 61	15 199
232	5 38 24	15.232
233	5 42 89	15.264
234	5 47 56	15.297
235	5 52 25	15 330
236	5 56 96	15 362
237	5 61 69	15.395
238	5 66 44	15.427
239	5 71 21	15 460
240	5 76 00	15 492
241	5 80 81	15 524
242	5 85 64	15.556
243	5 90 49	15 588
244	5 95 36	15 620
245	6 00 25	15.652
246	6 05 16	15.684
247	6 10 09	15.716
248	6 15 04	15.748
249	6 20 01	15 780
250	6 25 00	15.811
251	6 30 01	15 843
252	6 35 04	15.875
253	6 40 09	15.906
254	6 45 16	15.937
255	6 50 25	15.969
256	6 55 36	16.000
257	6 60 49	16.031
258	6 65 64	16.062
259	6 70 81	16 093
260	6 76 00	16.125
261	6 81 21	16.155
262	6 86 44	16.186
263	6 91 69	16 217
264	6 96 96	16.248
265	7 02 25	16 279
266	7 07 56	16.310
267	7 12 89	16 340
268	7 18 24	16 371
269	7 23 61	16.401
270	7 29 00	16 432
271	7 34 41	16.462
272	7 39 84	16.492
273	7 45 29	16 523
274	7 50 76	16.653
275	7 56 25	16.583
276	7 61 76	16 613
277	7 67 29	16 643
278	7 72 84	16.673
279	7 78 41	16.703
280	7 84 00	16.733
281	7 89 61	16 763
282	7 95 24	16.793
283	8 00 89	16.823
284	8 06 56	16.852
285	8 12 25	16.882
286	8 17 96	16.912
287	8 23 69	16.941
288	8 29 44	16 971
289	8 35 21	17.000
290	8 41 00	17.029
291	8 46 81	17.059
292	8 52 64	17.088
293	8 58 49	17.117
294	8 64 36	17 146
295	8 70 25	17.176
296	8 76 16	17 205
297	8 82 09	17.234
298	8 88 04	17.263
299	8 94 01	17 292
300	9 00 00	17.321

Squares and Square Roots of the Numbers from 1 to 1,000

Table V *(Continued)*

Number	*Square*	*Square Root*
501	25 10 01	22.383
502	25 20 04	22.405
503	25 30 09	22.428
504	25 40 16	22.450
505	25 50 25	22.472
506	25 60 36	22.494
507	25 70 49	22.517
508	25 80 64	22.539
509	25 90 81	22.561
510	26 01 00	22.583
511	26 11 21	22.605
512	26 21 44	22.627
513	26 31 69	22 650
514	26 41 96	22.672
515	26 52 25	22.694
516	26 62 56	22.716
517	26 72 89	22.738
518	26 83 24	22.760
519	26 93 61	22.782
520	27 04 00	22.804
521	27 14 41	22.825
522	27 24 84	22.847
523	27 35 29	22.869
524	27 45 76	22.891
525	27 56 25	22.913
526	27 66 76	22.935
527	27 77 29	22.956
528	27 87 84	22.978
529	27 98 41	23.000
530	28 09 00	23.022
531	28 19 61	23.043
532	28 30 24	23 065
533	28 40 89	23 087
534	28 51 56	23.108
535	28 62 25	23.130
536	28 72 96	23.152
537	28 83 69	23.173
538	28 94 44	23.195
539	29 05 21	23.216
540	29 16 00	23.238
541	29 26 81	23.259
542	29 37 64	23.281
543	29 48 49	23.302
544	29 59 36	23.324
545	29 70 25	23.345
546	29 81 16	23.367
547	29 92 09	23.388
548	30 03 04	23.409
549	30 14 01	23.431
550	30 25 00	23.452
551	30 36 01	23.473
552	30 47 04	23.495
553	30 58 09	23.516
554	30 69 16	23.537
555	30 80 25	23.558
556	30 91 36	23.580
557	31 02 49	23 601
558	31 13 64	23.622
559	31 24 81	23.643
560	31 36 00	23.664
561	31 47 21	23.685
562	31 58 44	23.707
563	31 69 69	23.728
564	31 80 96	23.749
565	31 92 25	23.770
566	32 03 56	23.791
567	32 14 89	23.812
568	32 26 24	23.833
569	32 37 61	23.854
570	32 49 00	23.875
571	32 60 41	23.896
572	32 71 84	23.917
573	32 83 29	23.937
574	32 94 76	23 958
575	33 06 25	23.979
576	33 17 76	24.000
577	33 29 29	24.021
578	33 40 84	24 042
579	33 52 41	24 062
580	33 64 00	24.083
581	33 75 61	24.104
582	33 87 24	24.125
583	33 98 89	24.145
584	34 10 56	24.166
585	34 22 25	24.187
586	34 33 96	24.207
587	34 45 69	24.228
588	34 57 44	24.249
589	34 69 21	24.269
590	34 81 00	24.290
591	34 92 81	24.310
592	35 04 64	24.331
593	35 16 49	24.352
594	35 28 36	24.372
595	35 40 25	24.393
596	35 52 16	24.413
597	35 64 09	24.434
598	35 76 04	24.454
599	35 88 01	24 474
600	36 00 00	24.495

Table V *Squares and Square Roots of the Numbers from 1 to 1,000 (Continued)*

Number	*Square*	*Square Root*	*Number*	*Square*	*Square Root*
401	16 08 01	20.025	451	20 34 01	21.237
402	16 16 04	20.050	452	20 43 04	21.260
403	16 24 09	20.075	453	20 62 09	21.284
404	16 32 16	20.100	454	20 61 16	21.307
405	16 40 25	20.125	455	20 70 25	21.331
406	16 48 36	20.149	456	20 79 36	21.354
407	16 56 49	20.174	457	20 88 49	21.378
408	16 64 64	20.199	458	20 97 64	21.401
409	16 72 81	20.224	459	21 06 81	21.424
410	16 81 00	20.248	460	21 16 00	21.448
411	16 89 21	20.273	461	21 25 21	21.471
412	16 97 44	20.298	462	21 34 44	21.494
413	17 05 69	20.322	463	21 43 69	21.517
414	17 13 96	20.347	464	21 52 96	21.541
415	17 22 25	20.372	465	21 62 25	21.564
416	17 30 56	20.396	466	21 71 56	21.587
417	17 38 89	20.421	467	21 80 89	21.610
418	17 47 24	20.445	468	21 90 24	21.633
419	17 55 61	20.469	469	21 99 61	21.656
420	17 64 00	20.494	470	22 09 00	21.679
421	17 72 41	20.518	471	22 18 41	21.703
422	17 80 84	20.543	472	22 27 84	21.726
423	17 89 29	20.567	473	22 37 29	21.749
424	17 97 76	20.591	474	22 46 76	21.772
426	18 06 25	20.616	475	22 56 25	21.794
426	18 14 76	20.640	476	22 65 76	21.817
427	18 23 29	20.664	477	22 75 29	21.840
428	18 31 84	20.688	478	22 84 84	21.863
429	18 40 41	20.712	479	22 94 41	21.886
430	18 49 00	20.736	480	23 04 00	21.909
431	18 67 61	20.761	481	23 13 61	21.932
432	18 66 24	20.785	482	23 23 24	21.954
433	18 74 89	20.809	483	23 32 89	21.977
434	18 83 56	20.833	484	23 42 56	22.000
435	18 92 25	20.857	485	23 52 25	22.023
436	19 00 96	20.881	486	23 61 96	22.045
437	19 09 69	20.905	487	23 71 69	22.068
438	19 18 44	20.928	488	23 81 44	22.091
439	19 27 21	20.952	489	23 91 21	22.113
440	19 36 00	20.976	490	24 01 00	22.136
441	19 44 81	21.000	491	24 10 81	22.159
442	19 53 64	21.024	492	24 20 64	22.181
443	19 62 49	21.048	493	24 30 49	22.204
444	19 71 36	21.071	494	24 40 36	22.226
445	19 80 25	21.095	495	24 50 25	22.249
446	19 89 16	21.119	496	24 60 16	22.271
447	19 98 09	21.142	497	24 70 09	22.293
448	20 07 04	21.166	498	24 80 04	22.316
449	20 16 01	21.190	499	24 90 01	22.338
450	20 25 00	21.213	500	25 00 00	22.361

Table V *Squares and Square Roots of the Numbers from 1 to 1,000 (Continued)*

Number	*Square*	*Square Root*	*Number*	*Square*	*Square Root*
701	49 14 01	26.476	751	56 40 01	27.404
702	49 28 04	26.495	752	56 55 04	27.423
703	49 42 09	26.514	753	56 70 09	27.441
704	49 56 16	26.533	754	56 85 16	27.459
705	49 70 25	26.552	755	57 00 25	27.477
706	49 84 36	26.571	756	57 15 36	27.495
707	49 98 49	26.589	757	57 30 49	27 514
708	50 12 64	26.608	758	57 45 64	27.532
709	50 26 81	26.627	759	57 60 81	27.550
710	50 41 00	26.646	760	57 76 00	27.568
711	50 55 21	26.665	761	57 91 21	27 586
712	50 69 44	26.683	762	58 66 44	27.604
713	50 83 69	26.702	763	58 21 69	27.622
714	50 97 96	26.721	764	58 36 96	27.641
715	51 12 25	26.739	765	58 52 25	27.659
716	51 26 56	26.758	766	58 67 56	27.677
717	51 40 89	26.777	767	58 82 89	27 695
718	51 55 24	26.796	768	58 98 24	27.713
719	51 69 61	26.814	769	59 13 61	27.731
720	51 84 00	26.833	770	59 29 00	27.749
721	51 98 41	26.851	771	59 44 41	27.767
722	52 12 84	26.870	772	59 59 84	27.785
723	52 27 29	26.889	773	59 75 29	27.803
724	52 41 76	26 907	774	59 90 76	27.821
725	52 56 25	26.926	775	60 00 25	27.839
726	52 70 76	26 944	776	60 21 76	27.857
727	52 85 29	26.963	777	60 37 29	27.875
728	52 99 84	26.981	778	60 52 84	27.893
729	53 14 41	27.000	779	60 68 41	27.911
730	53 29 00	27.019	780	60 84 00	27.928
731	53 43 61	27.037	781	60 99 61	27.946
732	53 58 24	27.055	782	61 15 24	27.964
733	53 72 89	27.074	783	61 30 89	27.982
734	53 87 56	27.092	784	61 46 56	28.000
735	54 02 25	27.111	785	61 62 25	28.018
736	54 16 96	27.129	786	61 77 96	28.036
737	54 31 69	27.148	787	61 93 69	28.054
738	54 46 44	27.166	788	62 09 44	28.071
739	54 61 21	27.185	789	62 25 21	28.089
740	54 76 00	27.203	790	62 41 00	28.107
741	54 90 81	27.221	791	62 56 81	28.125
742	55 05 64	27.240	792	62 72 64	28.142
743	55 20 49	27.258	793	62 88 49	28.160
744	55 35 36	27.276	794	63 04 36	28.178
745	55 50 25	27.295	795	63 20 25	28.196
746	55 65 16	27.313	796	63 36 16	28.213
747	55 80 09	27.331	797	63 52 09	28.231
748	55 95 04	27.350	798	63 68 04	28.249
749	56 10 01	27.368	799	63 84 01	28.267
750	56 25 00	27.386	800	64 00 00	28.284

Table V *Squares and Square Roots of the Numbers from 1 to 1,000* *(Continued)*

Number	*Square*	*Square Root*	*Number*	*Square*	*Square Root*
601	36 12 01	24.515	651	42 38 01	25.515
602	36 24 04	24.536	652	42 51 04	25.534
603	36 36 09	24.556	653	42 64 09	25.554
604	36 48 16	24.576	654	42 77 16	25 573
605	36 60 25	24.597	655	42 00 25	25.593
606	36 72 36	24.617	656	43 03 36	25.612
607	36 84 49	24.637	657	43 16 49	25 632
608	36 96 64	21 658	658	43 29 64	25.652
609	37 08 81	24.678	659	43 42 81	25 671
610	37 21 00	24.698	660	43 56 00	25.690
611	37 33 21	24.718	661	43 69 21	25.710
612	37 45 44	24.739	662	43 82 44	25.729
613	37 57 69	24.759	663	43 95 69	25.749
614	37 69 96	24.779	664	44 08 96	25.768
615	37 82 25	24.709	665	44 22 25	25.788
616	37 94 56	21 819	666	44 35 56	25 807
617	38 06 89	24.839	667	44 48 89	25 826
618	38 19 24	24.860	668	44 62 24	25.846
619	38 31 61	24 880	669	44 75 61	25.865
620	38 44 00	24.900	670	44 89 00	25.884
621	38 56 41	24.920	671	45 02 41	25.904
622	38 68 84	24.940	672	45 15 84	25 923
623	38 81 29	24.960	673	45 29 29	25 942
624	38 93 76	24.980	674	45 42 76	25 962
625	39 06 25	25.000	675	45 56 25	23.981
626	39 18 76	25.020	676	45 69 76	20 000
627	39 31 29	25 040	677	45 83 29	26.019
628	39 43 84	25.060	678	45 96 84	26 038
629	39 56 41	25.080	679	46 10 41	26.058
630	39 69 00	25.100	680	46 24 00	26.077
631	39 81 61	25.120	681	46 37 61	26 096
632	39 94 24	25.140	682	46 51 24	26.115
633	40 06 89	25.159	683	46 64 89	26.134
634	40 19 56	25.179	684	46 78 56	26.153
635	40 32 25	25.199	685	46 92 25	26.173
636	40 44 96	25 219	686	47 05 96	26.192
637	40 57 69	25.239	687	47 19 69	26 211
638	40 70 44	25 259	688	47 33 44	26 230
639	40 83 21	25 278	689	47 47 21	26.249
640	40 96 00	25.298	690	47 61 00	26 268
641	41 08 81	25.318	691	47 74 81	26 287
642	41 21 64	25 338	692	47 88 64	26 306
643	41 34 49	25 357	693	48 02 49	26 325
644	41 47 36	25 377	694	48 16 36	26 344
645	41 60 25	25 397	695	48 30 25	26 363
646	41 73 16	25 417	696	48 44 16	26.382
647	41 86 09	25 436	897	49 58 09	26 401
648	41 99 04	25 456	698	48 72 04	26 420
649	42 12 01	25 475	699	48 86 01	26 439
650	42 25 00	25.495	700	49 00 00	26 458

Squares and Square Roots of the Numbers from 1 to 1,000

Table V *(Continued)*

Number	*Square*	*Square Root*	*Number*	*Square*	*Square Root*
901	81 18 01	30.017	951	90 44 01	30.838
902	81 36 04	30 033	952	90 63 04	30.854
903	81 54 09	30.050	953	90 82 09	30 871
904	81 72 16	30 067	954	91 01 16	30.887
905	81 90 25	30.083	955	91 20 25	30.903
906	82 08 36	30.100	956	91 39 36	30.919
907	82 26 49	30.116	957	91 58 49	30.935
908	82 44 64	30.133	958	91 77 64	30.952
909	82 62 81	30.150	959	91 96 81	30.968
910	82 81 00	30.166	960	92 16 00	30.984
911	82 99 21	30.183	961	92 35 21	31.000
912	83 17 44	30.199	962	92 54 44	31.016
913	83 35 69	30.216	963	92 73 69	31.032
914	83 53 96	30.232	964	92 92 96	31.048
915	83 72 25	30.249	965	93 12 25	31.064
916	83 90 56	30.265	966	93 31 56	31.081
917	84 08 89	30.282	967	93 50 89	31.097
918	84 27 24	30.299	968	93 70 24	31.113
919	84 45 61	30.315	969	93 89 61	31.129
920	84 64 00	30.332	970	94 09 00	31.145
921	84 82 41	30.348	971	94 28 41	31.161
922	85 00 84	30 364	972	94 47 84	31.177
923	85 19 29	30.381	973	94 67 29	31.193
924	85 37 76	30.397	974	94 86 76	31.209
925	85 56 25	30.414	975	95 06 25	31.225
926	85 74 76	30.430	976	95 25 76	31.241
927	85 93 29	30.447	977	95 45 29	31.257
928	86 11 84	30.463	978	95 64 84	31.273
929	86 30 41	30.480	979	95 84 41	31.289
930	86 49 00	30.496	980	96 04 00	31.305
931	86 67 61	30.512	981	96 23 61	31.321
932	86 86 24	30.529	982	96 43 24	31.337
933	87 04 89	30.545	983	96 62 89	31.353
934	87 23 56	30.561	984	96 82 56	31.369
935	87 42 25	30.578	985	97 02 25	31.385
936	87 60 96	30.594	986	97 21 96	31.401
937	87 79 69	30.610	987	97 41 69	31.417
938	87 98 44	30 627	988	97 61 44	31.432
939	88 17 21	30.643	989	97 81 21	31.448
940	88 36 00	30.659	990	98 01 00	31.464
941	88 54 81	30 676	991	98 20 81	31.480
942	88 73 64	30 692	992	98 40 64	31.496
943	88 92 49	30.708	993	98 60 49	31.512
944	89 11 36	30.725	994	98 80 36	31.528
945	89 30 25	30.741	995	99 00 25	31.544
946	89 49 16	30.757	996	99 20 16	31.559
947	89 68 09	30.773	997	99 40 09	31.575
948	89 87 04	30.790	998	99 60 04	31.591
949	90 06 01	30 806	999	99 60 01	31.607
950	90 25 00	30.822	1000	100 00 00	31.623

Squares and Square Roots of the Numbers from 1 to 1,000

Table **V** *(Continued)*

Number	*Square*	*Square Root*	*Number*	*Square*	*Square Root*
801	64 16 01	28.302	851	72 42 01	29.172
802	64 32 01	28.320	652	72 59 04	29.189
803	64 48 09	28.337	853	72 76 09	29.206
804	64 64 16	28.355	854	72 93 16	29.223
805	64 80 25	28.373	855	73 10 25	29.240
806	64 96 36	28.390	856	73 27 36	29.257
807	65 12 49	28.408	857	73 44 49	29.275
808	65 28 64	28.425	858	73 61 64	29.292
809	65 44 81	28.443	859	73 78 81	29.309
810	65 61 00	28.460	860	73 96 00	29.326
811	65 77 21	28.478	861	74 13 21	29.343
812	65 93 44	28.496	862	74 30 44	29.360
813	66 09 69	28.513	863	74 47 69	29.377
814	66 25 96	28.531	864	74 64 96	29.394
815	66 42 25	28.548	865	74 82 25	29.411
816	66 58 56	28.566	866	74 99 56	29.428
817	66 74 89	28.583	867	75 16 89	29.445
818	60 91 24	28.601	868	75 34 24	29.462
819	67 07 61	28.618	869	75 51 61	29.479
820	67 24 00	28.636	870	75 69 00	29.496
821	67 40 41	28.653	871	75 86 41	29.513
822	67 56 84	28.671	872	76 03 84	29.530
823	67 73 29	28.688	873	76 21 29	29.547
824	67 89 76	28.705	874	76 38 76	29.563
825	68 00 25	28.723	875	76 56 25	29.580
826	68 22 76	28.740	876	76 73 76	29.597
827	68 39 29	28.758	877	76 91 29	29.614
828	68 55 64	28.775	878	77 08 84	29.631
829	68 72 41	28.792	879	77 26 41	29.648
830	68 89 00	28.810	880	77 44 00	29.665
831	69 05 61	28.827	881	77 61 61	29.682
832	69 22 24	28.844	882	77 79 24	29.698
833	69 38 89	28.862	883	77 96 89	29.715
834	69 55 56	28.879	884	78 14 56	29.732
835	69 72 25	28.896	885	78 32 25	29.749
836	69 88 96	28.914	886	78 49 96	29.766
837	70 05 69	28.931	887	78 67 69	29.783
838	70 22 44	28.948	888	78 85 44	29.799
839	70 39 21	28.965	889	79 03 21	29.816
840	70 56 00	28.983	890	79 21 00	29.833
841	70 72 81	29.000	891	79 38 81	29.850
842	70 89 64	29.017	892	79 56 64	29.866
843	71 06 49	29.034	893	79 74 49	29.883
844	71 23 36	29.052	894	79 92 36	29.900
845	71 40 25	29.069	895	80 10 25	29.916
846	71 57 16	29.086	896	80 28 16	29.933
847	71 74 09	29.103	897	80 46 09	29.950
848	71 91 04	29.120	898	80 64 04	29.967
849	72 08 01	29.138	899	80 82 01	29.983
850	72 25 00	29.155	900	81 00 00	30.000

Indexes

Index of Names

Index of Subjects